Geneious 4.8

Biomatters Ltd

March 4, 2010

Contents

1 Getting Started **7**

1.1 Downloading & Installing Geneious 7

1.2 Using Geneious for the first time 8

1.3 Troubleshooting 11

2 Retrieving and Storing data **15**

2.1 The main window 15

2.2 Importing and exporting data 23

2.3 Searching 31

2.4 Public databases 34

2.5 Storing data - Your Local Documents 40

2.6 Agents 45

2.7 Filtering and Similarity sorting 48

2.8 Notes 48

2.9 Preferences 52

2.10 Printing and Saving Images 56

3 Document Viewers **57**

3.1 The Sequence (and alignment) Viewer 57

3.2 Dotplot viewer 68

3.3 3D structure viewer 68

3.4 Tree viewer . . . 71
3.5 The Chromatogram viewer . . . 74
3.6 The PDF document viewer . . . 75
3.7 The Journal Article Viewer . . . 75

4 Analysing Data **77**
4.1 Literature . . . 77
4.2 Sequence data . . . 77
4.3 Dotplots . . . 78
4.4 Sequence Alignments . . . 79
4.5 Building Phylogenetic trees . . . 86
4.6 PCR Primers (*Pro* only) . . . 92
4.7 Contig Assembly (*Pro* only) . . . 101
4.8 Results of analysis . . . 108

5 Custom BLAST (*Pro* only) **109**
5.1 Setting Up . . . 109

6 COGs BLAST(*Pro* only) **113**
6.1 Setting Up . . . 113
6.2 BLASTing COGs . . . 114

7 Pfam (*Pro* only) **117**
7.1 Setting up the Pfam databases . . . 117
7.2 Pfam Document Types . . . 118
7.3 Pfam Operations . . . 119

8 Smart Folders (*Pro* only) **121**

9 Geneious Education (*Pro* only) **123**

9.1 Creating a tutorial . . . 123
9.2 Answering a tutorial . . . 124

10 Collaboration (*Pro* only) 125
10.1 Managing Your Accounts . . . 125
10.2 Managing Your Contacts . . . 128
10.3 Sharing Documents . . . 130
10.4 Browsing, Searching and Viewing Shared Documents . . . 130
10.5 Chat . . . 131

11 Cloning (*Pro* only) 133
11.1 Find Restriction Sites . . . 134
11.2 Digest into fragments . . . 135
11.3 Insert into Vector . . . 137
11.4 Gateway® Cloning . . . 140

12 Server Databases (*Pro* only) 143
12.1 Supported Database Systems . . . 143
12.2 Setting up . . . 144
12.3 Removing a server database . . . 145
12.4 Administration . . . 145

Chapter 1

Getting Started

One of the best ways to get an introduction to Geneious, its features and how to use them is to watch our online video demonstration: http://www.geneious.com/demonstration.

1.1 Downloading & Installing Geneious

Geneious is free to download from http://www.geneious.com/download. This download includes both Geneious Basic (free for academic use) and Geneious Pro. If you are using Geneious for the first time you will be offered a free trial of the Pro features. If you have already purchased a license you can enter it when Geneious starts up to activate the Pro features.

To download Geneious, click on the internet address above (or type it in to your internet browser) to open the Geneious download page then choose your operating system and click "Download Geneious". Geneious is available for Windows, Mac OS 10.4+, Linux and Solaris.

Once Geneious has downloaded, double left-click on the Geneious icon to start installing the program. While this is happening, you will be prompted for a location to install Geneious. Please check that you are satisfied with the location before continuing.

If you are using Mac OS X then you will only have to double click on the disk image that is downloaded then drag the Geneious application to your Applications folder.

1.1.1 Choosing where to store your data

When Geneious first starts up you will be asked to choose a location where Geneious will store all of your data. The default is normally fine but you may like to store your data on a network or USB drive so you can access it from other computers. To store your data on a different drive simply click the "Select" button in the welcome window and choose an empty folder on your

drive where you would like to store your data.

The data location can also be changed later by going to the "General" tab under "Tools" → "Preferences..." in the menu and changing the "Data Storage Location" option. Geneious will offer to copy your existing data across to the new location if appropriate.

1.1.2 Upgrading to new versions

To upgrade existing Geneious installations, simply download and install the new to the same location. This will retain all your data.

1.2 Using Geneious for the first time

Figure 1.1 shows the main Geneious window. This has six important areas or 'panels'.

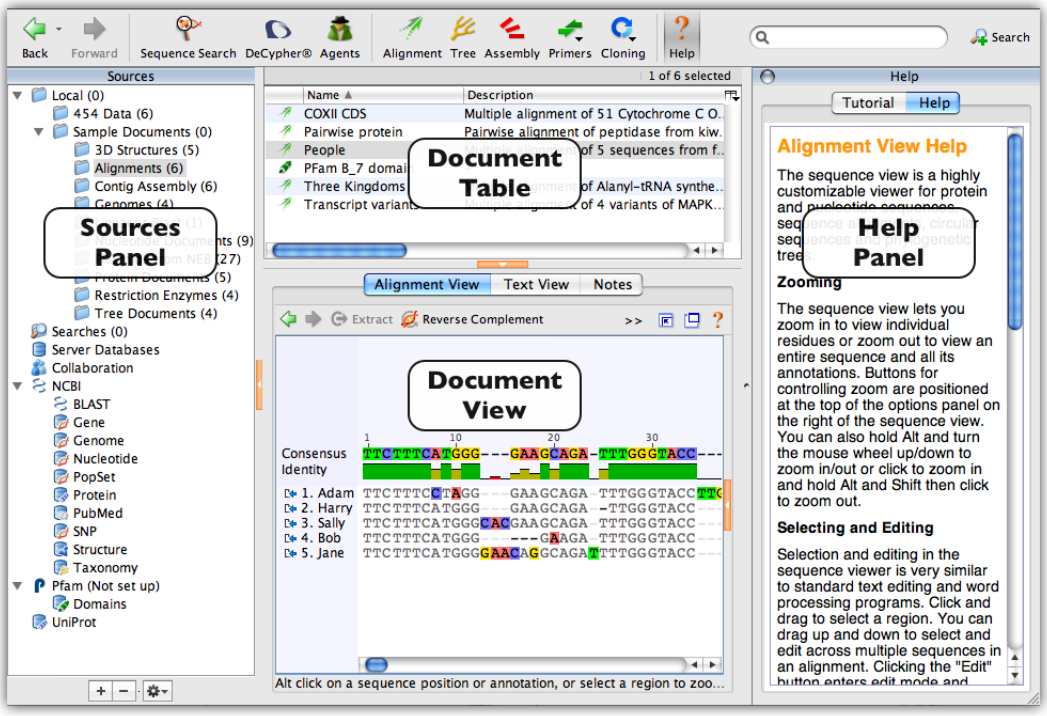


Figure 1.1: The main window in Geneious

1.2.1 The Sources Panel

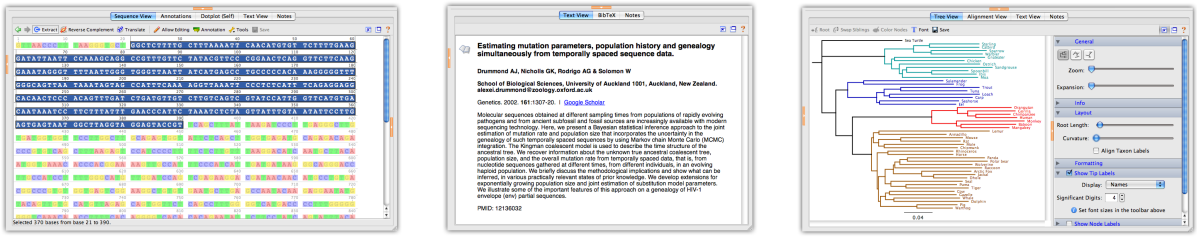
The Sources Panel contains the service Geneious offers for storing and retrieving data. These include your local documents (including sample documents), Server Databases, UniProt, NCBI, Pfam and Collaboration. All these services will be described in detail later in the manual. For more information see section 2.1.1.

1.2.2 The Document Table

The Document Table displays summaries of downloaded data such as DNA sequences, protein sequences, journal articles, sequence alignments, and trees. By clicking on the search icon you can search data for text or by sequence similarity (BLAST). You can enter a search string into the "Filter" box located at the right side of the toolbar; this will hide all documents that do not contain the search string. For more information, see section 2.1.2.

1.2.3 The Document Viewer Panel

The Document Viewer Panel is where sequences, alignments, trees, 3D structures, journal article abstracts and other types of documents can be shown graphically or as plain text. Many document viewers allow you to customize settings such as zoom level, color schemes, layout and annotations (nucleotide and amino acid sequences); three different layouts, branch and leaf labeling (tree documents); and many more. When viewing journal articles, this panel includes direct link to Google Scholar. All these options are displayed on the right-hand side of the panel (Figure 1.2). For more information see section 2.1.3


Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data.
Drummond AJ, Nicholls GK, Rodrigo AG & Solomon W

(a) Nucleotide sequence (b) Journal Article (c) Phylogenetic tree

Figure 1.2: Three document viewers

1.2.4 The Help Panel

The Help Panel has two sections: "Tutorial" and "Help". The tutorial gives you hands-on experience with some of the most popular features of Geneious. The Help section displays a short description of the currently selected service or document viewer. This panel can be closed at any time by clicking the button in its top corner, or by toggling the "Help" button in the Toolbar.

If you are new to Geneious, working through the tutorial is a great way to familiarize yourself with Geneious.

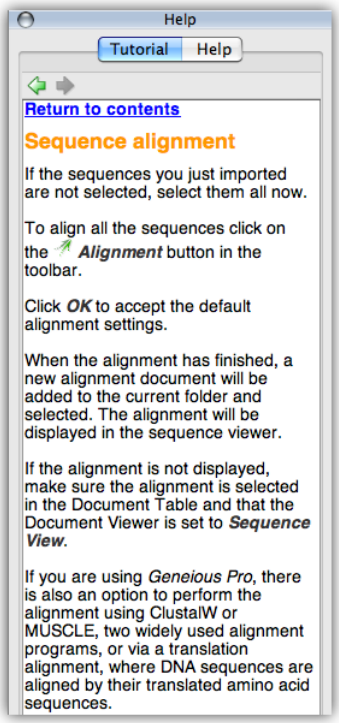

Help
Tutorial
Help
Return to contents
Sequence alignment
If the sequences you just imported are not selected, select them all now.
To align all the sequences click on the Alignment button in the toolbar.
Click OK to accept the default alignment settings.
When the alignment has finished, a new alignment document will be added to the current folder and selected. The alignment will be displayed in the sequence viewer.
If the alignment is not displayed, make sure the alignment is selected in the Document Table and that the Document Viewer is set to Sequence View.
If you are using Geneious Pro, there is also an option to perform the alignment using ClustalW or MUSCLE, two widely used alignment programs, or via a translation alignment, where DNA sequences are aligned by their translated amino acid sequences.

Figure 1.3: The Help Panel

1.2.5 The Toolbar

The toolbar gives quick access to commonly used features in Geneious including the *Search* for documents by keywords, *Agents* that search databases for new content even while you sleep, *Sort*ing sequences by similarity, pairwise or multiple sequence *Alignment*, *Tree* building, and *Help*. For more information on the toolbar, see section 2.1.5.

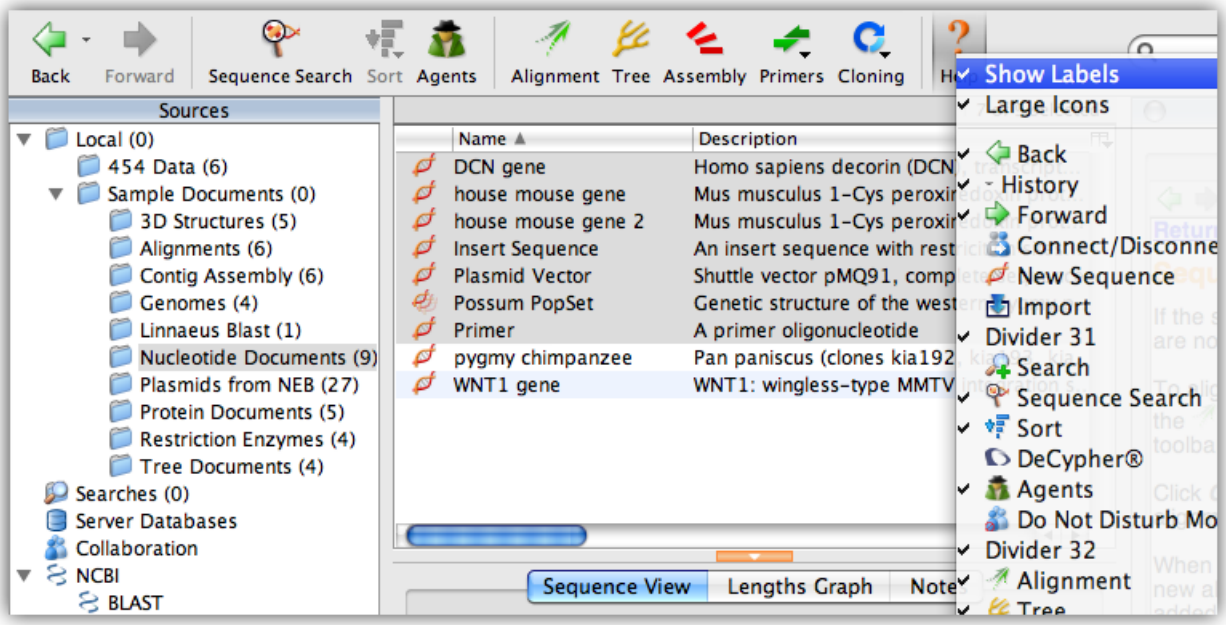


Figure 1.4: The Toolbar

1.2.6 The Menu Bar

The Menu Bar has seven main menus “File”, “Edit”, “View”, “Tools”, “Sequence”, “Collaboration”, “Help” and “Pro”. For details on the menu bar, see section 2.1.7.

1.2.7 Popup Menus

Many actions can be quickly accessed for data items, services and sometimes selections in a viewer via popup menus (also known as *context menus*). To invoke a popup menu for an item, simply right-click (Ctrl+click on Mac OS X). The popup menu will contain the actions which are relevant to the item you clicked.

1.3 Troubleshooting

1.3.1 Geneious won’t start

Geneious has some minimum system requirements. It is compatible with the three most common operating systems: Windows, Mac, and Linux. Check that you have one of the following

OS versions before you launch Geneious:

Operating System	System requirements
Windows	2000/XP/Vista
Mac OS	10.4/10.5
Unix/Linux	

Geneious also needs Java 1.5 to run. If you do not have this on your system already, please download a version of Geneious that includes Java 1.5 from http://www.geneious.com. This involves downloading a larger file.

If you are a Mac user, and have OS X 10.4 (Tiger), run Software Update and make sure you have the latest Java installed.

1.3.2 I get a connection error when trying to search using NCBI or EMBL

If the message reads, "Check your connection settings", there is a problem with your Internet connection. Make sure you are still connected to the Internet. Both Dial-up and Broadband can disconnect. If you are connected, then the error message indicates you are behind a proxy server and Geneious has been unable to detect you proxy settings automatically. You can fix this problem:

1. Check the browser you are using. These instructions are for Explorer, Safari, and Firefox.
2. Open your default browser.
3. Use the steps in Figure 1.5 for each browser to find the connection settings.
4. Now go into Geneious and select "Preferences". There are two ways to do this.
   - *Shortcut keys*. Ctrl+Shift+P (Windows/Linux), ⌘+Shift+P (Mac OS X).
   - *Tools Menu → Preferences*.
5. This opens the Preferences. Click on the "General" tab. There are five options in the drop-down options under "Connection settings" (Figure 1.6):
   - *Use direct connection*. Use this setting when no proxy settings are required.
   - *Use browser connection settings*. This allows Geneious to automatically import the proxy settings. This may not work with all web browsers.
   - *Use HTTP proxy server*. This enables two text fields : Proxy host and Proxy port. This information is in your browser's connection settings. Use this if your proxy server is an HTTP proxy server. Please see step 3.

- *Use SOCKS proxy server - Autodetect Type*. This enables two text fields : Proxy host and Proxy port. This information is in your browser's connection settings. Use this if your proxy server is a SOCKS proxy server. Please see step 3.
- *Use auto config file*. This enables one text field called "Config file location". These details can also be found in your browser's settings.

6. Set the proxy host and port settings under the General tab to match those in your browser.
7. If your proxy server requires a username and password you can specify these by clicking the "Proxy Password..." button directly below.

Note. If you are using any other browser, and cannot find the proxy settings, please email us at support@geneious.com.

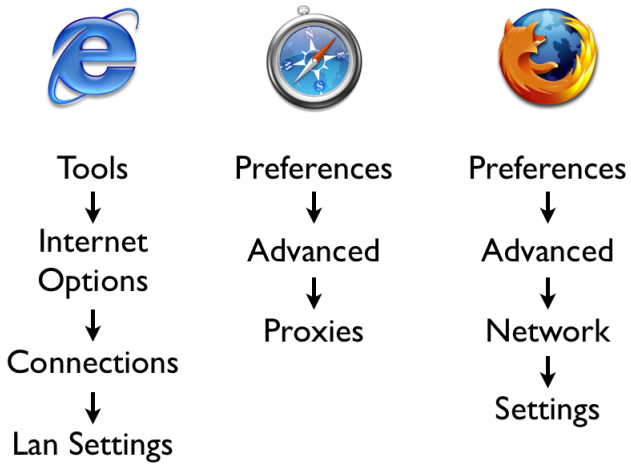


Figure 1.5: Checking browser settings

1.3.3 Web links inside Geneious don't work under Linux

Set your `BROWSER` environment variable to the name of your browser. The details depend on your browser and type of shell.

For example, if you are using Mozilla and bash, then put `export BROWSER=mozilla` in your `~/.bashrc` file. When using a csh shell variant, put `setenv BROWSER mozilla` in your `~/.cshrc` file.

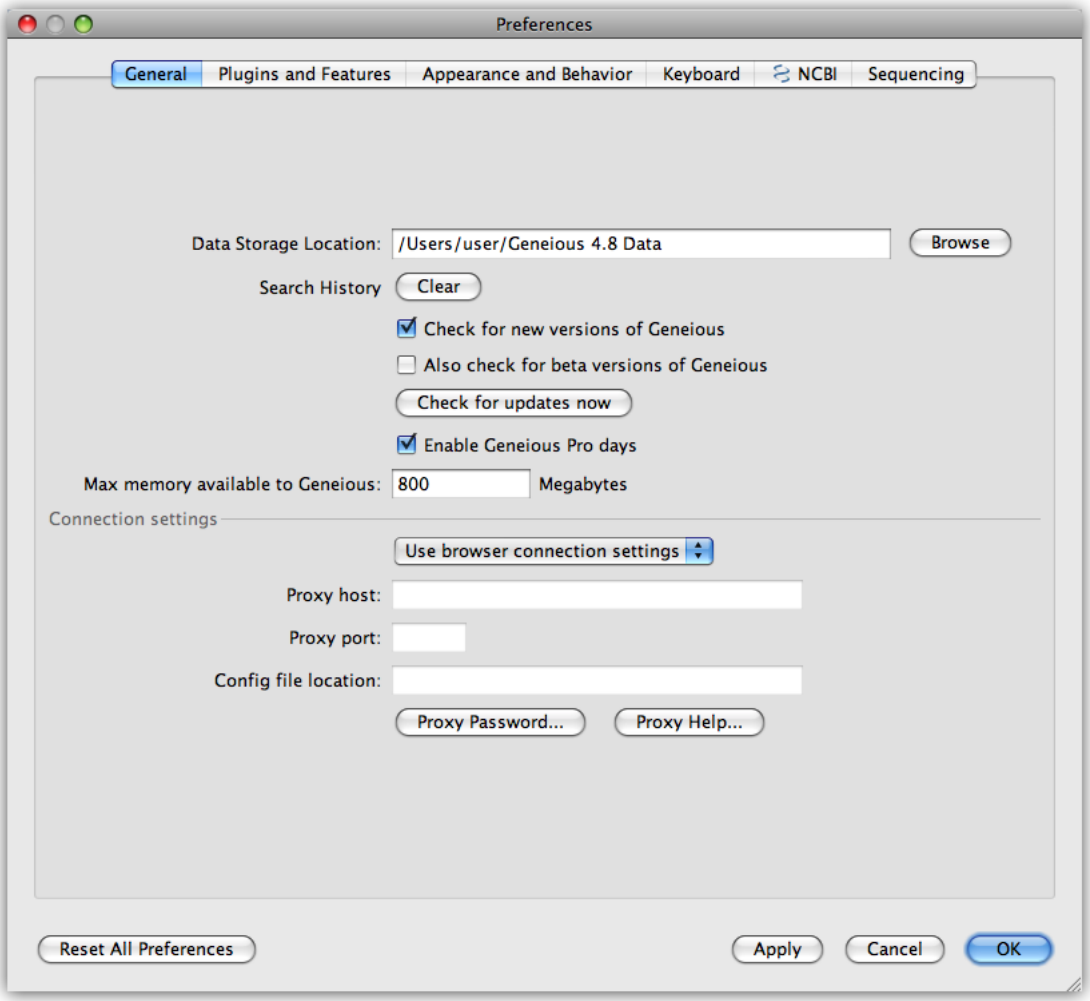


Figure 1.6: General Preferences

Chapter 2

Retrieving and Storing data

Geneious is a one-stop-shop for handling and managing your bioinformatic data. This chapter summarizes the different ways you can use Geneious to acquire, update, organize and store your data.

By the end of this chapter, you should be able to:

- Know the purpose of each panel in Geneious
- Import/Export data from various sources
- Organize your data into easily accessible folders
- Automatically update your data
- Know about the advantages of the "Note" functionality
- Customize Geneious to meet your needs.

2.1 The main window

This section provides more information on each of the panels in Geneious (Figure 2.1).

2.1.1 The Services Panel

The Services Panel shows a tree that concisely displays sources of data and your stored documents. The plus (+) symbol indicates that a folder contains sub-folders. A minus (-) indicates that the folder has been expanded, showing its sub-folders. Click these symbols to expand or contract folders.

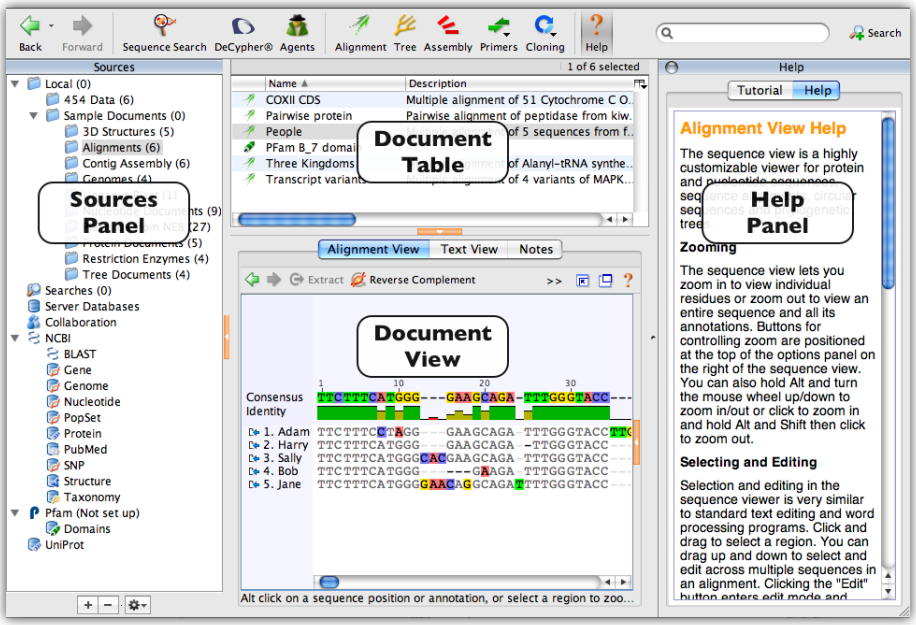


Figure 2.1: Geneious main window

Geneious Service Panel allows you to access:

- Your Local Documents.
- NCBI databases - Gene, Genome, Nucleotide, PopSet, Protein, Pubmed, SNP, Structure and Taxonomy.
- An EMBL database - Uniprot.
- Your contacts' Geneious databases.

You can view options for any selected service with the right mouse button, or by clicking the Options button at the bottom of the Service Panel in Mac OS X.

2.1.2 The Documents Table

The Document Table displays your search results or your stored documents. While search results usually contain documents of a single type, a local folder may contain any mixture of documents, whether they are sequences, publications or other types. If you cannot see all of the columns in the document table you may want to close the help panel to make more room.

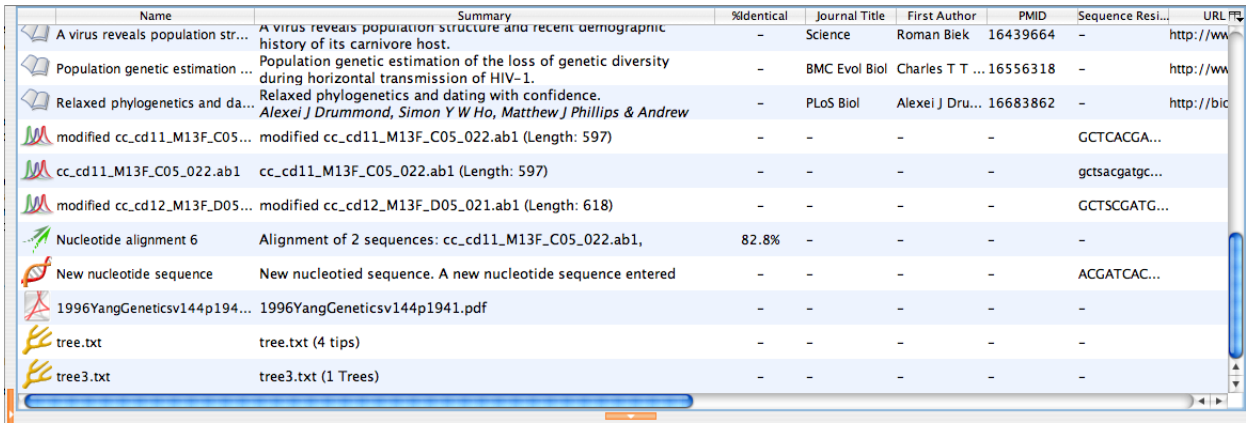


Figure 2.2: The document table, when browsing the local folders

This information is presented in table form (Figure 2.2).

Selecting a document in the Document Table will display its details in the Document View Panel. Selecting multiple documents will show a view of all the selected documents if they are of similar types. eg. Selecting two sequences will show both of them side-by-side in the sequence view. There are several ways to select multiple documents.

- Hold Ctrl (⌘ key on Mac OS X) and click to add the document to the current selection.
- Hold Shift and click to add the document and all documents between it and your previous selection.
- On windows the right mouse button can be clicked and held while moving the mouse to easily select a block of documents. The popup menu will appear once the mouse is released so the newly selected documents can quickly be manipulated.

Double-clicking a document in the Document Table displays the same view in a separate window.

To view the actions available for any particular document or group of documents, right-click (Ctrl+click on Mac OS X) on a selection of them. These options vary depending on the type of document.

The Document Table has some useful features.

Editing. Values can be typed into the columns of the table. This is a useful way of editing the information in a document. To edit a particular value, first click on the document and then click on the column which you want to edit. Enter the appropriate new information and press enter. Certain columns cannot be edited however, eg. the NCBI accession number.

Copying. Column values can be copied. This is a quick method of extracting searchable information such as an accession number. To copy a value, right-click (Ctrl+click on Mac OS X) on it, and choose the "Copy `name`" option, where `name` is the column name.

Sorting. All columns can be alphabetically, numerically or chronologically sorted, depending on the data type. To sort by a given column click on its header. If you have different types of documents in the same folder, click on the "Icon" column to sort then according to their type.

Managing Columns. You can reorder the columns to suit. Click on the column header and drag it to the desired horizontal position.

You can also choose which columns you want to be visible by right-clicking (Ctrl-click on Mac OS X) on any column header or by clicking the small header button in the top right corner of the table. This gives a popup menu with a list of all the available columns. Clicking on a column will show/hide it. Your preference is remembered so if you hide a column it will remain hidden in all areas of the program until you show it again.

As well as items to show/hide any of the available columns, there are a few more options in this popup menu to help you manage columns in Geneious.

- Lock Columns locks the state of the columns in the current table so that Geneious will never modify the way the columns are set up. You can still change the columns your self however.
- Save Column State... allows you to save the the current state if the columns so you can easily apply it to other tables. You can give the state a name and it will then appear in the Load Column State menu.
- Load Column State contains all of the columns states you have saved. Selecting a column state from here will immediately apply that state to the current table and lock the columns to maintain the new state. Use Delete Column State... to remove unwanted columns states from this menu.

Note. If a Note is added to a document (refer to the section on Adding Notes for more information), a Note column is added to the end of the existing Document table. Also, when accessing BLAST [1] in Geneious, the Document Table has additional columns related to the BLAST search.

2.1.3 The Document Viewer Panel

The Document Viewer Panel shows the contents of any document clicked on in the Document Table. To view large documents, it is sometimes better to double click on them. This opens a view in a new window. In the document viewer panel there are two tabs that are common to most types of documents: "Text view" and "Notes". "Text view" shows the document's

information in text format. The exception to this rule occurs with PDF documents where the user needs to either click the "View Document" button or double-click to view it.

Some document types such as sequences, trees and structures have an options panel occupying the right of the document viewer. The options in the options panel have an arrow which can be used to expand or hide a group of related options.

See the next section on document viewers for more information about operating the various viewers in Geneious.

Most viewers have their own small toolbar at the top of the document viewer panel. This always has three buttons on the far right:

- "Expand Document View" which expands the viewer panel out to fill the entire main window. Clicking again will return the viewer to normal size.
- "Open Document in New Window" will open a new view of the selected document in a new, separate window.
- "Help" opens the Help Panel and displays some short help for the current viewer.

2.1.4 The Help Panel

The Help Panel has a "Help" tab and a "Tutorial" tab.

The Help tab provides you information about the service you are currently using or the viewer you are currently viewing. The help displayed in the help tab changes as you click on different services and choose different viewers.

The Tutorial is aimed at first-time users of Geneious and has been included to provide a feel for how Geneious works. It is highly recommended that you work through the tutorial if you haven't used Geneious before.

2.1.5 The Toolbar

The toolbar contains several icons that provide shortcuts to common functions in Geneious. You can alter the contents of the toolbar to suit your own needs. The icons can be displayed small or large, and with or without their labels. The Help icon is always available.

The "Back" and "Forward" options help you move between previous views in Geneious and are analogous to the back and forward buttons in a web browser. The ▽ option shows a list of previous views. The other features that can be accessed from the toolbar are described in later sections.

The toolbar can be customized by right-clicking (Ctrl-click on Mac OS X) on it. This gives a popup menu with the following options:

- "Show Labels" Turn the text labels on or off.
- "Large Icons" Switch between large and small icons.
- A list of all available toolbar buttons. Selecting/deselecting buttons will show/hide the buttons in the toolbar.

2.1.6 Status bar

Below the Toolbar, there is a grey status bar. This bar displays the status of the currently selected service. For example, when you are running a search, it displays the number of matches, and the time remaining for the search to finish.

2.1.7 The Menu bar

File Menu

This contains some standard "File" menu items including printing and "Exit" on Windows. It also contains options to create, rename, delete, share and move folders and Import/Export options.

Edit Menu

Here you will find common editing functions including "Cut", "Copy", "Paste", "Delete" and "Select All". These are useful when transferring information from within documents to other locations, or exporting them. This menu also contains "Find in Document", "Find Next" and "Find Previous" options. Find can be used to find text or numbers in a selected document. This is useful when looking for annotated regions or a stretch of bases in a sequence. This opens a "Find Dialog". The shortcut to this is Ctrl+F. *Next* finds the next match for the text specified in the "Find" dialog. The shortcut keys are F3 or Ctrl+G. Geneious then allows you to choose another document and continue searching for the same search word. *Prev* finds the previous match. The shortcut keys for this are Ctrl+Shift+G or Shift+F3.

View Menu

This contains several options and commands for changing the way you view data in Geneious:

- "Back", "Forwards" and "History" allow you to return to documents you had selected previously.
- "Search" is discussed in section 2.3.
- "Agents" are discussed in section 2.6.
- "Next unread document" selects the next document in the current folder which is unread.
- "Table Columns" contains the same functionality as the popup menu for the document table header. See section 2.1.2 for more details.
- "Viewers" contains a list of available document viewers. Clicking one will select the view if it is available on the currently selected documents.
- "Open document in new window" Opens a new window with a view of the currently selected document(s).
- "Expand document view" expands the document viewer panel in the main window out to fill the entire main window. Selecting this again to return to normal.
- "Document Windows" Lists the currently open document windows. Selecting one from this menu will bring that document window to the front.

Tools Menu

- "Alignment" - see section 4.4
- "Tree" - see section 4.5
- "Assembly" - see section 4.7
- "Primers" - see section 4.6
- "Cloning" - see section 11
- "Sequence Search" - Perform a sequence search (such as NCBI Blast) using the currently selected sequence as the query. See section 2.4.4
- "Add/Remove Databases" - see section 5.1.3
- "Pfam" - see section 7
- "Linnaeus Blast" - Perform a blast search and display the results using the Linnaeus viewer. Evolutionary trees are built for hits within the same species. These are then displayed inside boxes nested according to the NCBI taxonomy.

- "Extract Annotations" - Search the selected sequences or alignments for annotations which match certain criteria then extract all of the matching annotations to separate sequence documents. Includes the option to concatenate all matches in each sequence into one sequence document. Useful for extracting a certain gene from a group of genomes.
- "Strip Alignment Columns" - creates a new alignment document with some columns (for example all identical columns or all columns containing only gaps) stripped
- "Concatenate Sequences or Alignments" - Joins the selected sequences or alignments end-on-end, creating a single sequence or alignment document from several. After selecting this operation you are given the option to choose the order in which the sequences or alignments are joined. You can also choose whether the resulting document is linear or circular, and, if one or more of the component sequences was an extraction from over the origin of a circular sequence, you can choose to use the numbering from that sequence, thus producing a circular sequence with its origin in the same place as the original circular sequence. Overhangs will be taken into account when concatenating.
- "Generate Consensus Sequence" - Generates a consensus sequence for the selected sequence alignment and saves it to a separate sequence document. After selecting this operation you are given options for choosing what type of consensus sequence you wish to generate - see section 3.1.5 for more details on the options.
- "Preferences" - see section 2.9

2.1.8 Sequence Menu

This contains several operations that can be performed on Protein and Nucleotide sequences as well as Sequence Alignments in some cases.

- *New Sequence* create a new nucleotide or protein sequence from residues that you can paste or type in.
- *Extract Region*, *Reverse Complement*, *Translate* see section 3.1.6 for details. Sometimes a selection in the sequence viewer is required before performing these.
- *Find ORFs...* Finds all open reading frames in a sequence and annotates them
- *Trim Ends...* See section 4.7.3.
- *Change Residue Numbering...* changes the "original residue numbering" of the selected sequence.
- *Convert between DNA and RNA* changes all T's in a sequence to U's or vice versa, depending on the type of the selected sequence. Once this is performed, click "Save" in the Sequence View to make the change permanent.

Collaboration Menu (*Pro* only)

This contains actions that can be performed with Collaboration accounts which allow you to share you work with other Geneious users.

Help Menu

This consists of the standard Help options offered by Geneious.

2.2 Importing and exporting data

Geneious is able to import raw data from different applications and export the results in a range of formats. If you are new to bioinformatics, please take the time to familiarize yourself with this chapter as there are a number of formats to be aware of.

2.2.1 Importing data from the hard drive to your Local folders

To import files from local disks or network drives, click "File" → "Import" → "From file". This will open up a file dialog. Select one or more files and click "Import". If Geneious automatic file format detection fails, select the file type you wish to import (Figure 2.3). The different file types are described in detail in the next section..

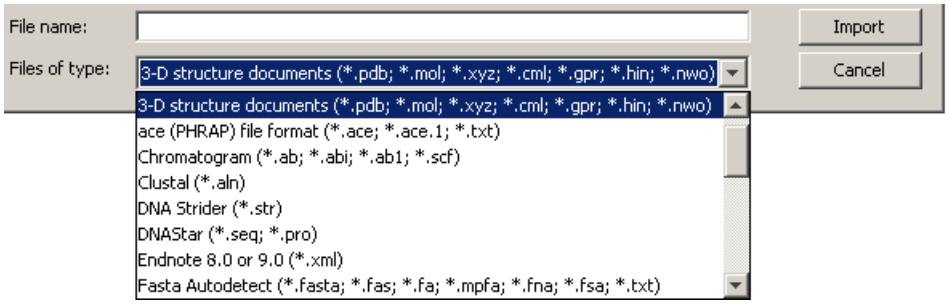

File name:
Import
Files of type:
3-D structure documents (*.pdb; *.mol; *.xyz; *.cml; *.gpr; *.hin; *.nwo)
Cancel
3-D structure documents (*.pdb; *.mol; *.xyz; *.cml; *.gpr; *.hin; *.nwo)
ace (PHRAP) file format (*.ace; *.ace.1; *.txt)
Chromatogram (*.ab; *.abi; *.ab1; *.scf)
Clustal (*.aln)
DNA Strider (*.str)
DNAStar (*.seq; *.pro)
Endnote 8.0 or 9.0 (*.xml)
Fasta Autodetect (*.fasta; *.fas; *.fa; *.mpfa; *.fna; *.fsa; *.txt)

Figure 2.3: File import options

2.2.2 Data input formats

Geneious version 4.8 can import the following file formats:

Format	Extensions	Data types	Common sources
Clustal	*.aln	Alignments	ClustalX
CSFASTA	*.csfasta	Color space FASTA	ABI SOLiD
DNAStar	*.seq, *.pro	Nucleotide & protein sequences	DNAStar
DNA Strider	*.str	Sequences	DNA Strider (Mac program), ApE
Embl/UniProt	*.embl, *.swp	Sequences	Embl, UniProt
Endnote (8.0) XML	*.xml	Journal article references	Endnote, Journal article websites
FASTA	*.fasta, *.fas, etc.	Sequences, alignments	PAUP*, ClustalX, BLAST, FASTA
FASTQ	*.fastq, *.fasq	Sequences with quality	Solexa/Illumina
GCG	*.seq	Sequences	GCG
GenBank	*.gb, *.xml	Nucleotide & protein sequences	GenBank
Geneious	*.xml, *.geneious	Preferences, databases	Geneious
Geneious Education	*.tutorial.zip	Tutorial, assignment etc.	Geneious
GFF	*.gff	Annotations	Sanger Artemis
MEGA	*.meg	Alignments	MEGA
Molecular structure	*.pdb'', *.mol, *.xyz, *.cml, *.gpr, *.hin, *.nwo	3D molecular structures	3D structure databases and programs
Newick	*.tre, *.tree, etc.	Phylogenetic trees	PHYLIP, Tree-Puzzle, PAUP*, ClustalX
Nexus	*.nxs, *.nex	Trees, Alignments	PAUP*, Mesquite, MrBayes & MacClade
PDB	*.pdb	3D Protein structures	SP3, SP2, SPARKS, Protein Data Bank
PDF	*.pdf	Documents, presentations	Adobe Writer, LaTeX, Miktex
Phrap ACE	*.ace	Contig assemblies	Phrap/Consed
PileUp	*.msf	Alignments	pileup (gcg)
PIR/NBRF	*.pir	Sequences, alignments	NBRF PIR
Qual	*.qual	Quality file	Associated with a FASTA file
Raw sequence text	*.seq	Sequences	Any file that contains only a sequence
Rich Sequence Format	*.rsf	Sequences, alignments	GCGs NetFetch
Sequence Chromatograms	*.ab1, *.scf	Raw sequencing trace & sequence	Sequencing machines
Vector NTI sequence	*.gb, *.gp	Nucleotide & protein sequences	Vector NTI
Vector NTI/AlignX alignment	*.apr	Alignments	Vector NTI, AlignX
Vector NTI Archive	*.ma4, *.pa4, *.oa4, *.ea4, *.ca6	Nucleotide & protein sequences, enzyme sets and publications	Vector NTI
Vector NTI/ContigExpress	*.cep	Nucleotide sequence assemblies	Vector NTI
Vector NTI database	VNTI Database	Nucleotide & protein sequences, enzyme sets and publications	Vector NTI

CLUSTAL format

The Clustal format is used by ClustalW [24] and ClustalX [23], two well known multiple sequence alignment programs.

Clustal format files are used to store multiple sequence alignments and contain the word clustal at the beginning. An example Clustal file:

```
CLUSTAL W (1.74) multiple sequence alignment

seq1 -----------------------KSKERYKDENGGNYFQLREDWWDANRETVWKAITCNA
```

```
seq2 ---------------YEGLTTANGXKEYYQDKNGGNFFKLREDWWTANRETVWKAITCGA
seq3 ----KRIYKKIFKEIHSGLSTKNGVKDRYQN-DGDNYFQLREDWWTANRSTVWKALTCSD
seq4 ------------------------SQRHYKD-DGGNYFQLREDWWTANRHTVWEAITCSA
seq5 --------------------NVAALKTRYEK-DGQNFYQLREDWWTANRATIWEAITCSA
seq6 ------FSKNIX--QIEELQDEWLLEARYKD--TDNYYELREHWWTENRHTVWEALTCEA
seq7 ------------------------------------------------KELWEALTCSR

seq1 --GGGKYFRNTCDG--GQNPTETQNNCRCIG----------ATVPTYFDYVPQYLRWSDE
seq2 P-GDASYFHATCDSGDGRGGAQAPHKCRCDG---------ANVVPTYFDYVPQFLRWPEE
seq3 KLSNASYFRATC--SDGQSGAQANNYCRCNGDKPDDDKP-NTDPPTYFDYVPQYLRWSEE
seq4 DKGNA-YFRRTCNSADGKSQSQARNQCRC---KDENGKN-ADQVPTYFDYVPQYLRWSEE
seq5 DKGNA-YFRATCNSADGKSQSQARNQCRC---KDENGXN-ADQVPTYFDYVPQYLRWSEE
seq6 P-GNAQYFRNACS----EGKTATKGKCRCISGDP----------PTYFDYVPQYLRWSEE
seq7 P-KGANYFVYKLD-----RPKFSSDRCGHNYNGDP---------LTNLDYVPQYLRWSDE
```

CSFASTA format

ABI .csfasta files represent the color calls generated by the SOLiD sequencing system.

DNAStar files

DNAStar .seq and .pro files are used in Lasergene, a sequence analysis tool produced by DNAStar.

DNA Strider

Sequence files generated by the Mac program DNA Strider, containing one Nucleotide or Protein sequence.

EMBL/UniProt

Nucleotide sequences from the EMBL Nucleotide Sequence Database, and protein sequences from UniProt (the Universal Protein Resource)

EndNote 8.0 XML format

EndNote is a popular reference and bibliography manager. EndNote lets you search for journal articles online, import citations, perform searches on your own notes, and insert references into documents. It also generates a bibliography in different styles. Geneious can interoperate with

EndNote using Endnote's XML (Extensible Markup Language) file format to export and import its files.

FASTA format

The FASTA file format is commonly used by many programs and tools, including BLAST [1], T-Coffee [17] and ClustalX [23]. Each sequence in a FASTA file has a header line beginning with a ">" followed by a number of lines containing the raw protein or DNA sequence data. The sequence data may span multiple lines and these sequence may contain gap characters. An empty line may or may not separate consecutive sequences. Here is an example of three sequences in FASTA format (DNA, Protein, Aligned DNA):

```
>Orangutan
ATGGCTTGTGGTCTGGTCGCCAGCAACCTGAATCTCAAACCTGGAGAGTGCCTTCGAGTG

>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
QIWQK

>Chicken
CTACCCCCCTAAAACACTTTGAAGCCTGATCCTCACTA------------------CTGT
CATCTTAA
```

FASTQ format

FASTQ format stores sequences and Phred qualities in a single file.

GenBank files

Records retrieved from the NCBI webiste (`http://www.ncbi.nlm.nih.gov`) can be saved in a number of formats. Records saved in GenBank or INSDSeq XML formats can be imported into Geneious.

Geneious format

The Geneious format can be used to store all your local documents, note types and program preferences. A file in Geneious format will usually have a `.geneious` extension or a `.xml` extension. This format is useful for sharing documents with other Geneious users and backing up your Geneious data.

Geneious Education format

This is an archive containing a whole bundle of files which together comprise a Geneious education document. This format can be used to create assignments for your students, bioinformatics tutorials, and much more. See chapter 9 for information on how to create such files.

GFF format

The GFF format contains sequence annotation information (and optional sequences). You can use a GFF file to annotate existing sequences in your local database, import entirely new sequences, or import the annotations onto blank sequences.

MEGA format

The MEGA format is used by MEGA (Molecular Evolutionary Genetics Analysis).

Molecular structure

Geneious imports a range of molecular structure formats. These formats support showing the locations of the atoms in a molecule in 3D:

- **PDB format** files from the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Database
- ***.mol format** files produced by MDL Information Systems Inc
- ***.xyz format** files produced by XMol
- ***.cml format** files in Chemical Markup Language
- ***.gpr format** ghemical files
- ***.hin format** files produced by HyperChem
- ***.nwo format** files produced by NWChem

Newick format

The Newick format is commonly used to represent phylogenetic trees (such as those inferred from multiple sequence alignments). Newick trees use pairs of parentheses to group related taxa, separated by a comma (,). Some trees include numbers (branch lengths) that indicate the

distance on the evolutionary tree from that taxa to its most recent ancestor. If these branch lengths are present they are prefixed with a colon (:). The Newick format is produced by programs such as PHYLIP, PAUP*, ClustalW [24], ClustalX [23], Tree-Puzzle [8] and PROTML. Geneious is also able to read trees in Newick format and display them in the visualization window. It also gives you a number of display options including tree types, branch lengths, and labels.

Nexus format

The Nexus format [13] was designed to standardize the exchange of phylogenetic data, including sequences, trees, distance matrices and so on. The format is composed of a number of blocks such as TAXA, TREES and CHARACTERS. Each block contains pre-defined fields. Geneious imports and exports files in Nexus format, and can process the information stored in them for analysis.

PDB format

Protein Databank files contain a list of XYZ co-ordinates that describe the position of atoms in a protein. These are then used to generate a 3D model which is usually viewed with Rasmol or SPDB viewer. Geneious can read PDB format files and display an interactive 3D view of the protein structure, including support for displaying the protein's secondary structure when the appropriate information is available.

PDF format

PDF stands for Portable Document Format and is developed and distributed by Adobe Systems (`http://www.adobe.com/`). It contains the entire description of a document including text, fonts, graphics, colors, links and images. The advantage of PDF files is that they look the same regardless of the software used to create them. Some word processors are able to export a document into PDF format. Alternatively, Adobe Writer can be used. Currently, you can use Geneious to read, store and open PDF files and future versions will have more options for storing and manipulating PDF.

Phrap Ace files

Ace is the format used by the Phrap/Consed package, created by the University of Washington Genome Center. This package is used mainly to assemble sequences.

PileUp format

The PileUp format is used by the pileup program, a part of the Genetics Computer Group (GCG) Wisconsin Package.

PIR/NBRF format

Format used by the Protein Information Resource, a database established by the National Biomedical Research Foundation

Qual file

Quality file which must be in the same folder as the sequence file (FASTA format) for the quality scores to be used.

Raw sequence format

A file containing only a sequence

Rich Sequence format

RSF (Rich Sequence Format) files contain one or more sequences that may or may not be related. In addition to the sequence data, each sequence can be annotated with descriptive sequence information.

Sequence Chromatograms

Sequence chromatogram documents contain the results of a sequencing run (the trace) and a guess at the sequence data (base calling).

Informally, the trace is a graph showing the concentration of each nucleotide against sequence positions. Base calling software detects peaks in the four traces and assigns the most probable base at more or less even intervals.

Vector NTI® formats

Geneious supports the import of several Vector NTI formats:

- ***.gb and *.gp formats** These formats are used in Vector NTI for saving single nucleotide and protein sequence documents. They are very similar to the GenBank formats with the same extensions, although they contain some extra information.
- ***.apr format** This format is used for storing alignments and trees made with AlignX, Vector NTI's alignment module.
- ***.ma4, *.pa4, *.oa4, *.ea4 and *.ca6 formats** These are the archive formats which Vector NTI uses to export whole databases.
- ***.cep format** This format is produced by the ContigExpress module and Geneious will import sequences (including the positions of the base calls), traces, qualities, trimmed regions, annotations and editing history for individual reads and contigs.

2.2.3 Where does my imported data go?

The above formats can be all imported into Geneious from local files. Geneious also enables you to download certain types of documents directly from public databases such as NCBI and EMBL. The method used to retrieve a particular piece of data will determine where in Geneious it is stored.

Data imported from local files. This is imported directly into the currently selected local folder within Geneious. If no folder is selected, Geneious will open a dialog which lets you specify a folder.

Data from an NCBI/EMBL/Contacts search. Data downloaded from public databases within Geneious will appear in the Document Table when that database is selected and can be dragged from there into a local folder of your choice.

Important: if you don't drag the documents from a database search into your local folders the results will be lost when Geneious is closed.

2.2.4 Data output formats

Each data type has several export options. Any set of documents may be exported in Geneious native format.

Additionally, documents imported in any chromatogram or molecular structure format can be re-exported in that format as long as no changes have been made to the document.

2.2.5 Export to comma separated values (CSV) file

The value displayed in the document table can be exported to csv file which can be loaded by most spread sheet programs. When choosing to export in csv format Geneious will also present

Data type	Export format options
DNA sequence	FASTA, Genbank XML, Genbank flat, Geneious
Amino acid sequence	FASTA, Genbank XML, Genbank flat, Geneious
Protein 3D structure	PDB, FASTA, Geneious
Multiple sequence alignment	Phylip (`*.phy`), FASTA, NEXUS [13], MEGA3 [12], Geneious
Phylogenetic tree	Phylip (`*.phy`), FASTA, NEXUS [13], Newick, MEGA3 [12], Geneious
PDF document	PDF, Geneious
Publication	EndNote 8.0, Geneious

a list of the available columns in the table (including hidden ones) so you can choose which to export. Please note this format cannot be imported currently.

2.2.6 Batch Export

Batch export takes the selected documents and exports each to its own file. Eg. Select several chromatograms to export them all to ab1 format files. The options for batch export let you specify the format and folder to export to as well as the extension to use. Each file will be named according to the Name column in Geneious.

2.3 Searching

Searching is designed to be as user-friendly as possible and the process is the same if you are searching your local documents or a public database such as NCBI. To search the selected database or folder click the "Search" button from the toolbar. For non-local folders search will be on by default and cannot be closed. This applies to NCBI and EMBL databases. For local folders search is off by default.

When search is first activated the document table will be emptied to indicate no results have been found. To return to browsing click the "Search" button again or press the Escape key while the cursor is in the search text field.

To initiate a search enter the desired search term(s) in the text field and press enter or click the adjacent "Search" button. Once a search starts the results will appear in the document table as they are found. The "Search" button changes to a "Cancel" button while a search is in progress and this may be clicked at any time to terminate the search. Feedback on a search progress is presented in the status bar directly below the toolbar (see Figure 2.4).

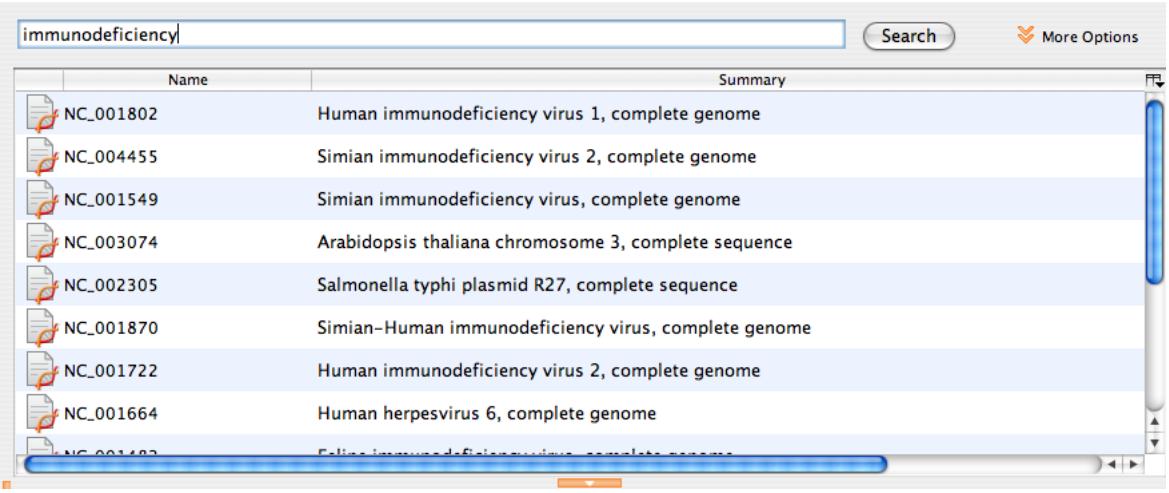


Figure 2.4: The Search tab of the Document Table

2.3.1 Advanced Search options

To access advanced search click the "More Options" button inside the basic search panel. To return to basic search click the "Fewer Options" button. Switching between advanced and basic will not clear the search results table.

This feature provides more search options to select from. Geneious allows you to search with a range of criteria; however, these depend on the database being searched. All the fields in the NCBI public databases can be searched in any combination. Each database has a specific list of fields and it is important to familiarize yourself with these fields to make full use of the Advanced Search. The fields available for a search can be found in the left-most drop-down box after enabling the advanced search options.

Note. When searching the Genome database, the documents returned are only summaries. To download the whole genome, select the summary(s) of the genome(s) you would like to download and the click the "Download" button inside the document view or just above it. There are also "Download" items in the File menu and in the popup menu when document summary is right-clicked (Ctrl+click on Mac OS X). The size of these files is not displayed in the Documents Table. Be aware that whole genomes can be very large and can take a long time to download. You can cancel the download of document summaries by selecting "Cancel Downloads" from any of the locations mentioned above.

Advanced Search also provides you with a number of options for restricting the search on a field depending on the field you are searching against. For example, if you are using numbers to search for "Sequence length" or "No. of nodes" you can further restrict your search with the second drop-down box:

- "is greater than" (>)
- "is less than" (<)
- "is greater than or equal to" (≥)
- "is less than or equal to" (≤)

Likewise if you are searching on the "Creation Date" search field you have the following options

- "is before or on"
- "is after or on"
- "is between"

When searching your local folders you have the option of searching by "Document type". The second drop-down list provides the options "is" and "is not". The third drop-down lists the various types of documents that can be stored in Geneious such as "3D-Structure", "Nucleotide sequence", and "PDF" (see Figure 2.5).

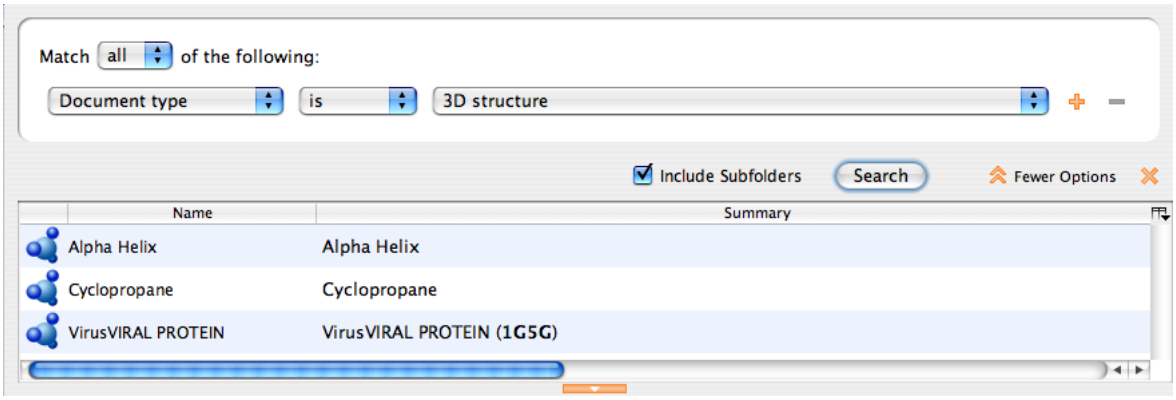


Figure 2.5: Document type search options

And/Or searches

The advanced options lets you search using multiple criteria. By clicking the "+" button on right of the search term you can add another search criteria. You can remove search criteria by clicking on the appropriate "-" button. The "Match all/any of the following" option at the top of the search terms determines how these criteria are combined:

Match "Any" requires a match of one or more of your search criteria. This is a broad search and results in more matches.

Match "All" requires a match all of your search criteria. This is a narrow search and results in fewer matches.

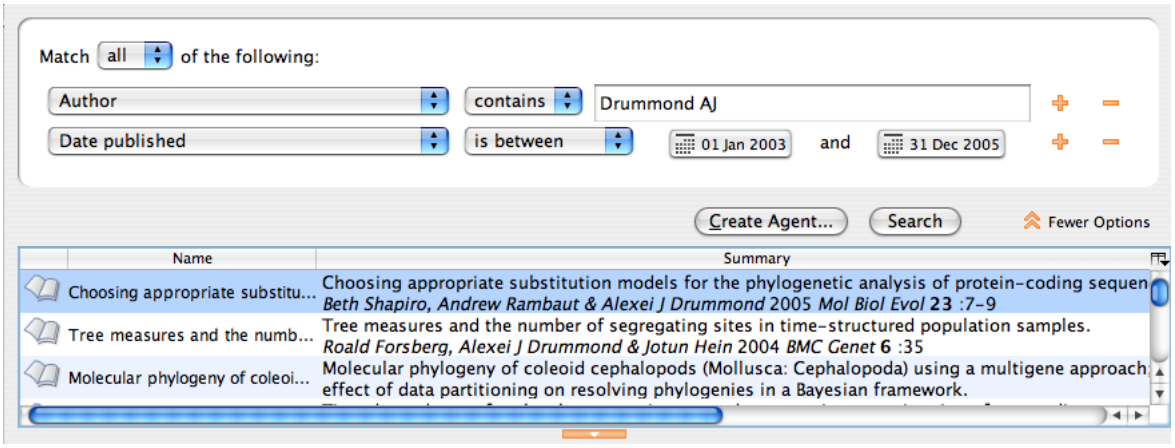

Match all of the following:
Author
contains
Drummond AJ
Date published
is between
01 Jan 2003
and
31 Dec 2005
Create Agent...
Search
Fewer Options
Name
Summary
Choosing appropriate substitu...
Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequen
Beth Shapiro, Andrew Rambaut & Alexei J Drummond 2005 Mol Biol Evol 23 :7-9
Tree measures and the numb...
Tree measures and the number of segregating sites in time-structured population samples.
Roald Forsberg, Alexei J Drummond & Jotun Hein 2004 BMC Genet 6 :35
Molecular phylogeny of coleoi...
Molecular phylogeny of coleoid cephalopods (Mollusca: Cephalopoda) using a multigene approach
effect of data partitioning on resolving phylogenies in a Bayesian framework.

Figure 2.6: Advanced Search

2.3.2 Autocompletion of search words

Geneious remembers previously searched keywords and offers an auto-complete option. This works in a similar way to Google or predictive text on your mobile phone. If you click within the search field, a drop-down box will appear showing previously used options.

2.4 Public databases

Geneious allows you to search several public databases in the same way that you can search your local documents. The search process is described in section 2.3.

Geneious is able to communicate with a number of public databases hosted by the National Centre for Biotechnology Information (NCBI) as well as the UniProt and Pfam databases. You can access these databases through the web at `http://www.ncbi.nlm.nih.gov`, `http://www.uniprot.org/` and `http://www.sanger.ac.uk/Software/Pfam/` respectively. These are all well known and widely used storehouses of molecular biology data.

When viewing data from a public database such as NCBI the data can not be modified. This is demonstrated by the small padlock icon which appears in the status bar. When this icon is

present items cannot be added or removed from the table and they cannot be modified in any way. To modify an item you must first move it to your local folders.

2.4.1 Pfam

See chapter 7.

2.4.2 UniProt

This database is a comprehensive catalogue of protein data. It includes protein sequences and functions from Swiss-Prot, TrEMBL, and PIR. It has three main components, each optimized for a particular purpose.

2.4.3 NCBI (Entrez) databases

NCBI was established in 1988 as a public resource for information on molecular biology. Geneious allows you to directly download information from nine important NCBI databases and perform NCBI BLAST searches (Table 2.1).

Table 2.1: NCBI databases accessible via Geneious

Database	Coverage
Genome	Whole genome sequences
Nucleotide	DNA sequences
PopSet	sets of DNA sequences from population studies
Protein	Protein sequences
Structure	3D structural data
PubMed	Biomedical literature citations and abstracts
Taxonomy	Names and taxonomy of organisms
SNP	Single Nucleotide Polymorphisms
Gene	Genes

The Entrez Genome database. This provides views of a variety of genomes, complete chromosomes, sequence maps with contigs (contiguous sequences), and integrated genetic and physical maps.

The Entrez Nucleotide database. This database in GenBank contains 3 separate components that are also searchable databases: "EST", "GSS" and "CoreNucleotide". The core nucleotide database brings together information from three other databases: GenBank, EMBI, and DDBJ. These are

part of the International collaboration of Sequence Databases. This database also contains RefSeq records, which are NCBI-curated, non-redundant sets of sequences.

The Entrez Popset database. This database contains sets of aligned sequences that are the result of population, phylogenetic, or mutation studies. These alignments usually describe evolution and population variation. The PopSet database contains both nucleotide and protein sequence data, and can be used to analyze the evolutionary relatedness of a population.

The Entrez Protein database. This database contains sequence data from the translated coding regions from DNA sequences in GenBank, EMBL, and DDBJ as well as protein sequences submitted to the Protein Information Resource (PIR), SWISS-PROT, Protein Research Foundation (PRF), and Protein Data Bank (PDB) (sequences from solved structures).

The Entrez Structure database. This is NCBI's structure database and is also called MMDB (Molecular Modeling Database). It contains three-dimensional, biomolecular, experimentally or programmatically determined structures obtained from the Protein Data Bank.

The PubMed database. This is a service of the U.S. National Library of Medicine that includes over 16 million citations from MEDLINE and other life science journals. This archive of biomedical articles dates back to the 1950s. PubMed includes links to full text articles and other related resources, with the exception of those journals that need licenses to access their most recent issues.

Entrez Taxonomy. This database contains the names of all organisms that are represented in the NCBI genetic database. Each organism must be represented by at least one nucleotide or protein sequence.

Entrez Gene. Entrez Gene is NCBI's database for gene-specific information. It does not include all known or predicted genes; instead Entrez Gene focuses on the genomes that have been completely sequenced, that have an active research community to contribute gene-specific information, or that are scheduled for intense sequence analysis.

Entrez SNP. In collaboration with the National Human Genome Research Institute, The National Center for Biotechnology Information has established the dbSNP database to serve as a central repository for both single base nucleotide subsitutions and short deletion and insertion polymorphisms.

The scope and depth of these databases make them critical information sources for molecular biologists and bioinformaticians alike. However, a library is only as good as its librarian. Geneious is your librarian, allowing you to search for, filter and store, only the data that you care about.

2.4.4 Accessing NCBI BLAST through Geneious

BLAST [1] stands for Basic Local Alignment Search Tool. It allows you to query the NCBI sequence databases with a sequence in order to find entries in the public database that contain

similar sequences. When "BLAST-ing", you are able to specify either nucleotide or protein sequences and nucleotide sequences can be either DNA or RNA sequences. The result of a BLAST query is a table of "hits". Each hit refers to a GenBank accession number and the gene or protein name of the sequence. Each hit also has a "Bit-score" which provides information about how similar the hit is to the query sequence. The bigger the bit score, the better the match. Finally there is also an "E-value" or "Expect value", which represents the number of hits with at least this score that you would expect purely by chance, given the size of the database and query sequence. The lower the E-value, the more likely that the hit is real.

Geneious can perform seven different kinds of BLAST search:

- **blastn**: Compares a nucleotide query sequence against a nucleotide sequence database.
- **Megablast**: A variation on blastn that is faster but only finds matches with high similarity.
- **Discontiguous Megablast**: A variation on blastn that is slower but more sensitive. It will find more dissimilar matches so it is ideal for cross-species comparison.
- **blastp**: Compares an amino acid query sequence against a protein sequence database.
- **blastx**: Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
- **tblastn**: Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
- **tblastx**: Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Please note that the tblastx program cannot be used with the nr database on the BLAST Web page because it is too computationally intensive.

Geneious is able to run NCBI BLAST on many different databases. Some of these databases are non-redundant in order to reduce duplicate hits. You can submit either a raw sequence or Genbank accession number into NCBI BLAST and receive a summary of results for each hit. This summary contains the bit-score, e-value, identity, and the stretch of the query sequence and hit sequence that match. The databases that can be searched are shown in the following tables.

You can quickly and easily BLAST against any of these databases using any of the available BLAST programs via the Sequence Search operation. This operation can be accessed by going to the Tools menu or by right-clicking (Ctrl+click on Mac OS X) on a sequence document and choosing "Sequence Search". This will bring up the sequence search options.

Geneious gives you the option of searching against a database using either your currently selected sequence documents or a sequence you enter manually. If you choose to enter your

Table 2.2: Nucleotide sequence searches in the BLAST databases

Database	Nucleotide searches
nr	All non-redundant GenBank+EMBL+DDBJ+PDB sequences(no EST, STS, GSS or HTGS sequences)
genome	Genomic entries from NCBI's Reference Sequence project
est	Database of GenBank + EMBL + DDBJ sequences from EST Divisions
est_human	Human subset of est
est_mouse	Mouse subset of est
est_others	Non-Human, non-mouse subset of est
gss	Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
htgs	Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2 (finished, phase 3 HTG sequences are in nr)
pat	Nucleotide sequences derived from the Patent division of GenBank
PDB	Sequences derived from the 3D-structures of proteins from PDB
month	All new/updated GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
RefSeq	NCBI-curated, non-redundant sets of sequences.
dbsts	Database of GenBank+EMBL+DDBJ sequences from STS Divisions
chromosome	A database with complete genomes and chromosomes from the NCBI Reference Sequence project.
wgs	A database for whole genome shotgun sequence entries.
env_nt	This contains DNA sequences from the environment, i.e all organisms put together

Table 2.3: Protein sequence searches in the BLAST databases

Database	Protein searches
env_nr	Translations of sequences in env_nt
month	All new/updated GenBank coding region (CDS) translations +PDB+SwissProt+PIR released in last 30 days
nr	All non-redundant GenBank coding region (CDS) translations+PDB+SwissProt+PIR+PRF
pat	Protein sequences derived from the Patent division of GenBank
PDB	Sequences derived from 3D structure Brookhaven PDB
RefSeq	RefSeq protein sequences from NCBI's Reference Sequence Project
SwissProt	Curated protein sequences information from EMBL

sequence manually, then Geneious will display a large text box in which you can enter your query sequence as either unformatted text or FASTA format.

Select your database using the first drop-down box. Databases are grouped together under their respective services. The available programs in the second drop-down box will depend on the database you have chosen.

Geneious also allows you to specify most of the advanced options that are available in BLAST. To access the advanced options click the "More Options" button which is in the bottom left of the Sequence Search options. The available options vary depending on the kind of BLAST search you have selected. For details on each of the options you can hover your mouse over the option to see a short description or refer to the NCBI BLAST documentation at http://www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml.

Once a search has started, a results folder will be created under the Searches folder in the Sources panel. Search progress is shown in the document table. The search can be cancelled by

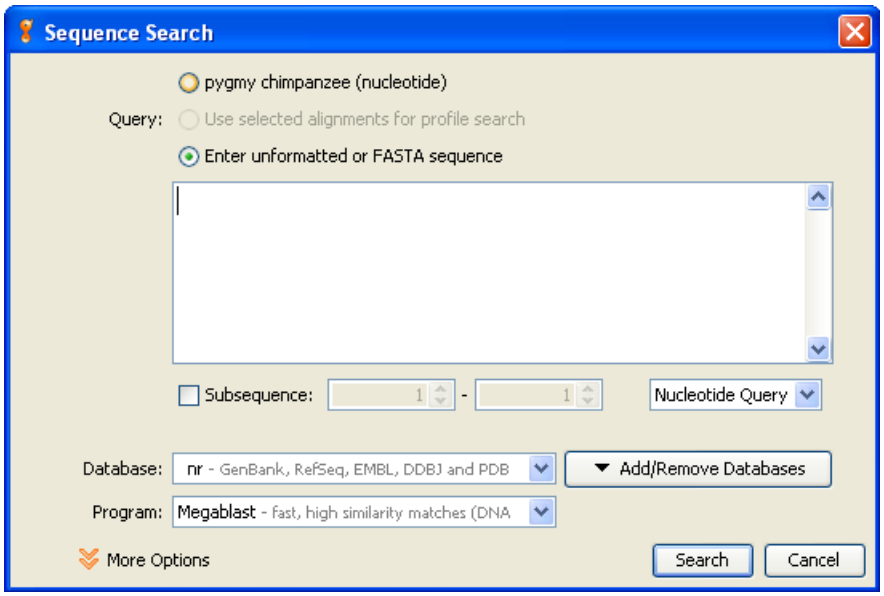


Figure 2.7: Sequence Search Options

clicking on the red square labelled "Stop" (See Figure 2.8).

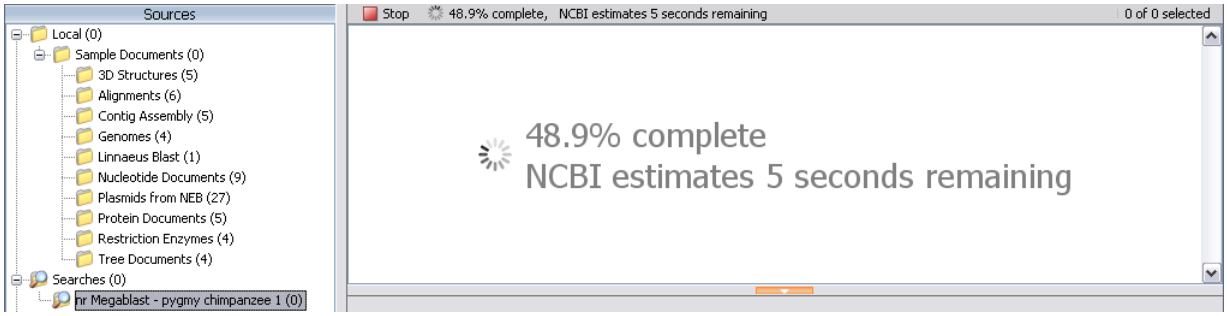


Figure 2.8: Sequence Search in Progress

Once the search has completed the results can be moved to your local database at your convenience. If your query sequence was annotated then any annotations that cover the hit region will be transfered to the BLAST hit document.

You can also download the full database sequence that corresponds to a BLAST hit. To retrieve the full sequence select a BLAST alignment and go to $File \rightarrow Download\ Documents$ or click the **Download Full Sequence(s)** button located above the viewer tabs. The full sequence will be available in the "Sequence View" tab once the download has completed. In addition the annotations from the full sequence will be transfered over to the BLAST alignment (see Figure 2.10).

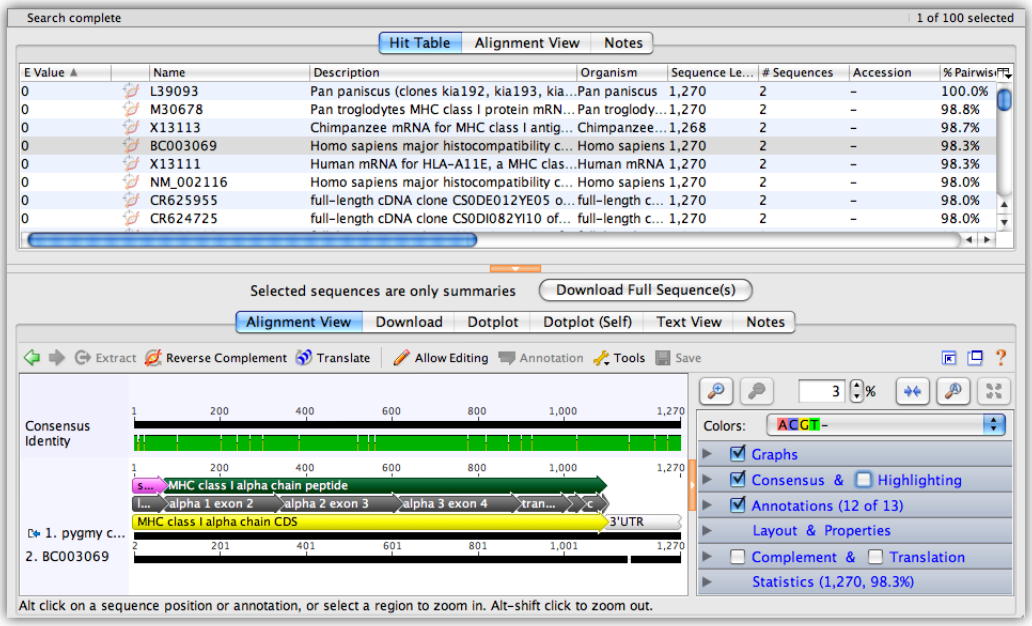

Search complete
1 of 100 selected
Hit Table
Alignment View
Notes
E Value
Name
Description
Organism
Sequence Le...
Sequences
Accession
% Pairwis
0 L39093 Pan paniscus (clones kia192, kia193, kia... Pan paniscus 1,270 2 - 100.0%
0 M30678 Pan troglodytes MHC class I protein mRN... Pan troglody... 1,270 2 - 98.8%
0 X13113 Chimpanzee mRNA for MHC class I antig... Chimpanzee... 1,268 2 - 98.7%
0 BC003069 Homo sapiens major histocompatibility c... Homo sapiens 1,270 2 - 98.3%
0 X13111 Human mRNA for HLA-A11E, a MHC clas... Human mRNA 1,270 2 - 98.3%
0 NM_002116 Homo sapiens major histocompatibility c... Homo sapiens 1,270 2 - 98.0%
0 CR625955 full-length cDNA clone CS0DE012YE05 o... full-length c... 1,270 2 - 98.0%
0 CR624725 full-length cDNA clone CS0DI082YI10 of... full-length c... 1,270 2 - 98.0%
Selected sequences are only summaries
Download Full Sequence(s)
Alignment View
Download
Dotplot
Dotplot (Self)
Text View
Notes
Extract
Reverse Complement
Translate
Allow Editing
Annotation
Tools
Save
Consensus Identity
MHC class I alpha chain peptide
alpha 1 exon 2
alpha 2 exon 3
alpha 3 exon 4
tran...
3'UTR
MHC class I alpha chain CDS
1. pygmy c...
2. BC003069
Colors:
Graphs
Consensus & Highlighting
Annotations (12 of 13)
Layout & Properties
Complement & Translation
Statistics (1,270, 98.3%)
Alt click on a sequence position or annotation, or select a region to zoom in. Alt-shift click to zoom out.

Figure 2.9: Sequence Search Complete

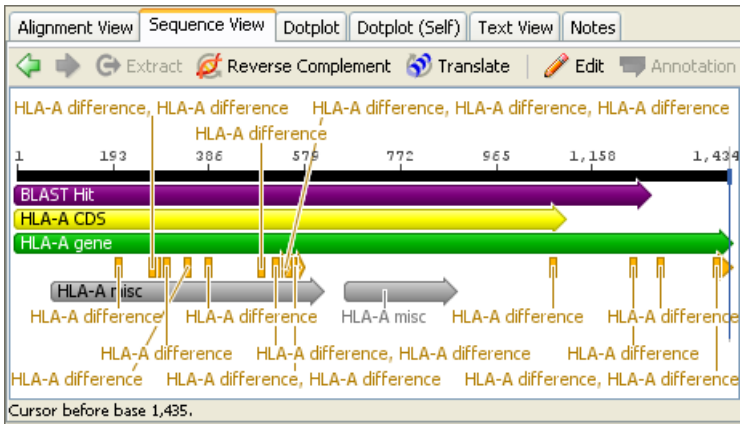
If you have a mirror of the NCBI BLAST databases you can set Geneious to use this by going to *Tools* → *Add/Remove Databases* → *Set Up Search Services*. This will bring up a dialog that allows you to change the setup for various search services in Geneious. Choose NCBI using the service drop-down box at the top of the dialog. Enter the url for the mirror and click OK to apply the new settings. You can also edit the databases that show up in Geneious by clicking on Edit Databases. This will only change the databases that Geneious displays and will not have any effect on the actual databases on the BLAST server.

2.5 Storing data - Your Local Documents

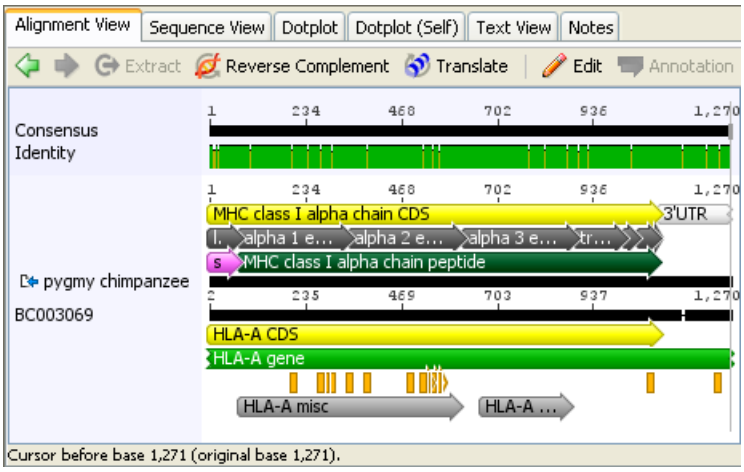
Geneious can be used to store your documents locally. Under the "Local" folder in the Services Panel you are able to create sub-folders to organize and store a variety of document types (Table 2.4).

This is also where you can set up special folders to receive documents that are downloaded by a Geneious agent. To create a new folder in Geneious, select the "Local" folder or a sub-folder icon in the services panel and right-click (Ctrl+click on Mac OS X). This will pop up a menu. Clicking on "New folder..." opens a dialog that will prompt you to name the folder. The named folder is then created as a sub-folder of the folder that you originally right-clicked on.

Important. Search results will be lost when you exit Geneious unless the downloaded docu-



(a) Complete Database Sequence



(b) Annotations Transfered to Alignment

Figure 2.10: Document After Full Sequence Download

Table 2.4: Geneious document types

Document type	Geneious Icon
Nucleotide sequence	
Protein sequence	
Phylogenetic tree	
3D structure	
Sequence alignment	
Chromatogram	
Journal articles	
PDF	
Other documents	

ments have been copied or moved to one of your local folders.

In Geneious you can create new folders, rename existing folders, delete and export folders. All these choices are available by either right-clicking on the folder, clicking on the action menu (Mac OS X), or by holding down the Ctrl button and clicking (Mac OS X). Also in Mac OS X, you can also use the plus (+) and minus (-) buttons located at the bottom of the service panel to create and delete folders.

2.5.1 Transferring data

It is quick and easy to transfer data to your local folders from either a Geneious database search or from your computer's hard drive. Please check you have already set up your destination folders before continuing.

Moving documents from Geneious searches to your Local folders

There are a number of ways to do this.

Drag and drop. This is quickest and easiest. Select the documents that you want to move. Then, while holding the mouse button down, drag them over to the desired folder and release. If you dragged documents from one local folder to another, this action will move the documents – so that a copy of the document is not left in the original location. In external databases such as NCBI the documents will be copied, leaving one in its original location.

Drag and copy. While dragging a document over to your folder, hold the Ctrl key (Alt/Option key on Mac OS X) down. This places a copy of the document in the target folder while leaving a copy in the original location. This is useful if you want copies in different folders. Folders themselves can also be dragged and dropped to move them but they cannot be copied.

The Edit menu. Select the document and then open the Edit menu on the menu bar. Click on "Cut" (Ctrl+X/⌘+X), or "Copy" (Ctrl+C/⌘+C). Select the destination folder and "Paste" (Ctrl+V/⌘+V) the document into it.

2.5.2 Searching your Local folders

The "Services Panel" allows you to browse your Local folder hierarchy. Next to each folder name in the hierarchy is the number of documents it contains in brackets. When the Local folder or a sub-folder is collapsed (minimized), the brackets next to the folder shows how many files are contained in that folder as well as all of its sub-folders. In addition, if some of the documents in a/ folder are unread, the number of unread documents will also appear in the brackets.

You can search the Local folder (and sub-folders) the same way you search the public databases by clicking on the "Search" icon. If you have defined a new type of note in Geneious, and a Note has been added, it will also be added to the "Advanced Search" criteria. Look at an example of a new Note type called "Protein size", which takes a text value for the protein in kDa (kiloDaltons) (see Figure 2.11).

Important: You must use quotation marks ("") if "!", "@", "$", and blank spaces (" ") are part of your search criteria. No quotation marks lead to unreliable results.

Wild card searches

When you are looking for all matches to a partial word, use the asterisk (*). For example, typing "oxi*" would return matches such as oxidase, oxidation, oxido-reductase, and oxide. This is useful for performing generic searches. You can also place the asterisk (*) in the middle of the word but not at the beginning. This feature is available only for local documents.

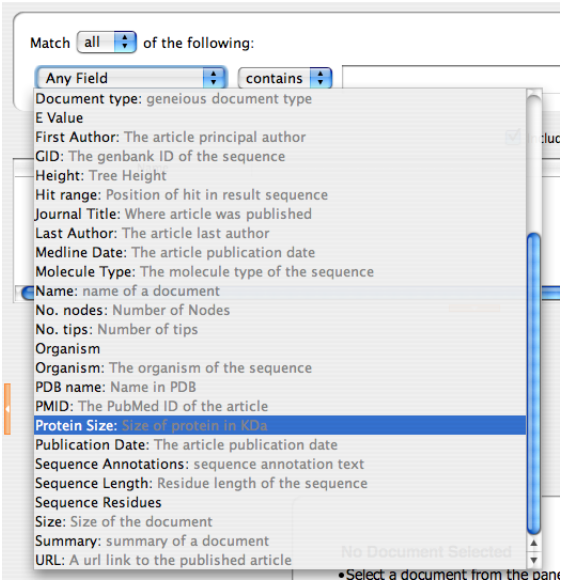

Match all of the following:
Any Field
contains
Document type: geneious document type
E Value
First Author: The article principal author
GID: The genbank ID of the sequence
Height: Tree Height
Hit range: Position of hit in result sequence
Journal Title: Where article was published
Last Author: The article last author
Medline Date: The article publication date
Molecule Type: The molecule type of the sequence
Name: name of a document
No. nodes: Number of Nodes
No. tips: Number of tips
Organism
Organism: The organism of the sequence
PDB name: Name in PDB
PMID: The PubMed ID of the article
Protein Size: Size of protein in KDa
Publication Date: The article publication date
Sequence Annotations: sequence annotation text
Sequence Length: Residue length of the sequence
Sequence Residues
Size: Size of the document
Summary: summary of a document
URL: A url link to the published article
No Document Selected
•Select a document from the ban

Figure 2.11: Searching the local documents on a user-defined field

Similarity ("BLAST-like") searching

It is possible to search your local documents not only for text occurrences but by similarity to sequence fragments. Click the small arrow at the bottom of the large T to the left of the search dialog, select "Nucleotide similarity search" or "Protein similarity search" and enter the sequence text. Geneious will try to guess the type of search based on the text, so that simply entering or pasting a sequence fragment may change the search type automatically.

The search locates documents containing a similar string of residues, and orders them in decreasing order of similarity to the string. The ordering is based on calculating an e-value for each match. You can read more about the e-value in subsection 2.4.4.

For the search to be successful, you need to specify a minimum of 11 nucleotides and 3 amino acids. Note that search times depend on the number and size of your sequence documents, and so may take a long time to complete.

2.5.3 Checking and changing the location of your Local folders

To check where your Local folders are being stored on your hard drive, open the Tools menu in the Menu Bar. Click "Tools" → "Preferences" → "General". Your documents are stored at the location specified by the "Data Storage Location" field (see Figure 2.12). You can change this location by clicking the "Browse" button and selecting a new location. Geneious will remember this new location when you exit.

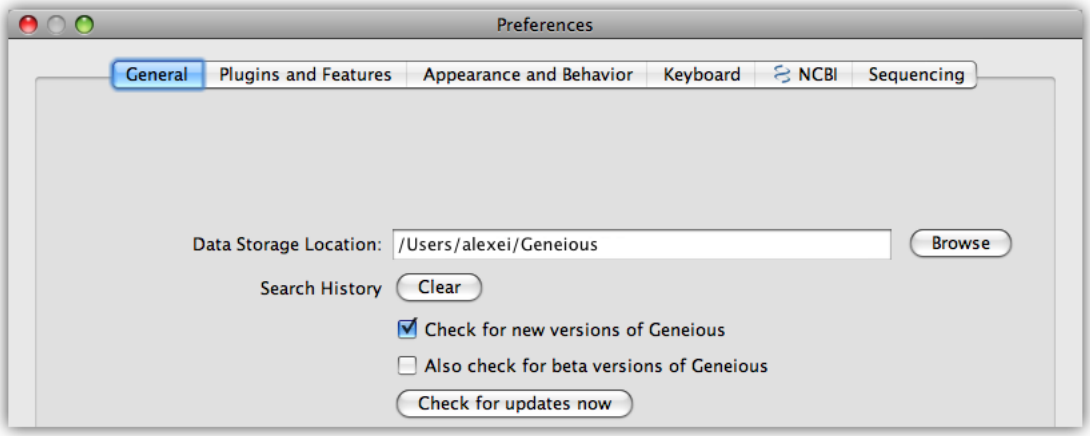


Figure 2.12: Setting the location of your local documents

2.6 Agents

Geneious offers a simple way for you to continuously receive the latest information on genomes, sequences, and protein structures. This feature is called an agent. Each agent is a user-defined, automated search. You can instruct an agent to search any Geneious accessible database at regular intervals (eg. weekly) including your contacts on Collaboration. This simple but powerful feature ensures that you never miss that critical article or DNA sequence. To manage agents click on the agent icon in the toolbar. An agent has to be set up before it can be used.

2.6.1 Creating agents

To set up an Agent click the Agents icon and the create button. You now need to specify a set of search criteria in the exact same way as you do for search, the database to search, search frequency and the folder you wish the agent to deliver its results to.

The search frequency may be specified in minutes, hours, days or weeks. You can only use whole numbers.

Selecting "Only get documents created after today" will cause the agent to check what documents are currently available when the agent is created. Then when the agent searches it will only get documents that are new since it was created. e.g. If you have already read all publications by a particular author and you want the agent to only get publications released in the future.

Alternatively you can click the "Create Agent..." button which is available in some advanced search panels. This will use the advanced search options you have entered to create the agent.

The easiest way to organize your search results is to create a new folder and name it appropriately. You can do that by navigating to the parent folder in the "Deliver to" box and click "New Folder", or by creating a new folder beforehand,

1. Right-click (Ctrl+click on Mac OS X) on the "Sample Documents" or "Local" folders. This brings up a popup menu with a "New Folder..." option.

2. Create a new folder and name it according to the contents of the search. (For example, type "CytB" if searching for cytochrome b complex.)

3. Once created, select the new folder. You can now select the "Create" or "Create and Run". The agent will then be added to the list in the agent dialog and it will perform its first search if you clicked "Create and Run". Otherwise it will wait until its next scheduled search.

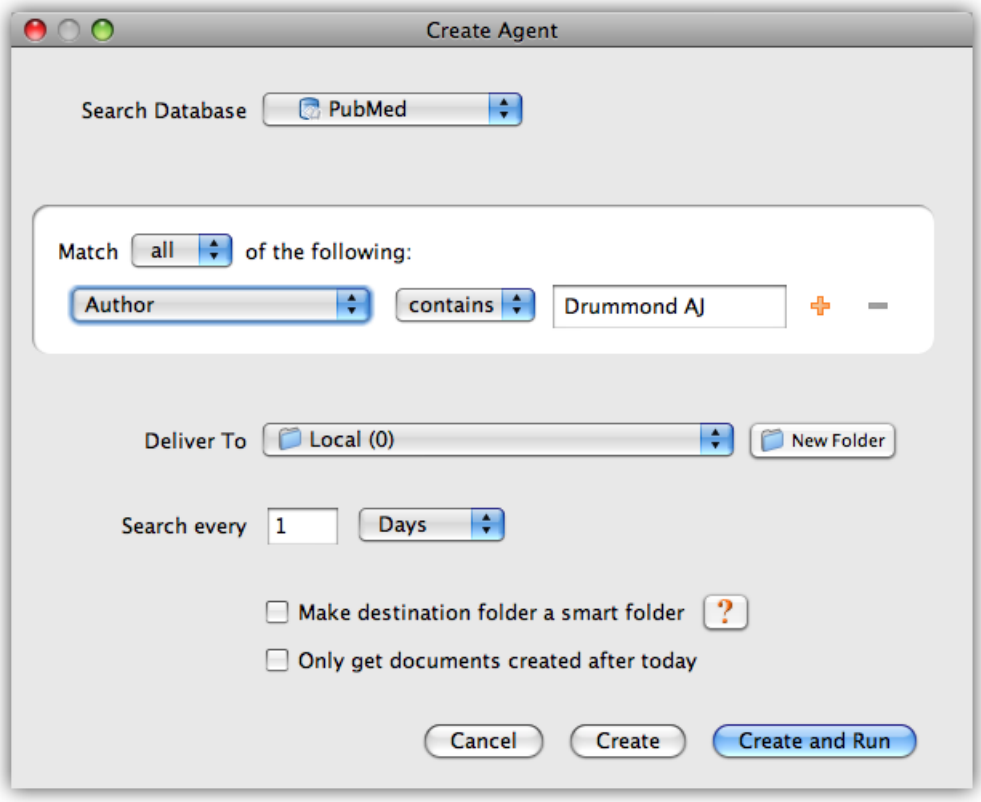


Figure 2.13: The Create Agent Dialog

2.6.2 Checking agents

Once you have created one or more agents, Geneious allows you to quickly view their status in the agents window which is accessible from the toolbar. Your agents' details are presented in several columns: *Enable*, *Action*, *Status* and *Deliver To*.

Enable This column contains a check box showing whether the agent is enabled. *Action*. This summarizes the user-defined search criteria. It contains:

1. Details of the database accessed. For example, Nucleotide and Genome under NCBI.
2. The search type the Agent performed, i.e. "keyword".
3. The words the user entered in the search field for the Agent to match against.

Status. This indicates what the Agent is currently doing. The status will be one of the following:

- "Next search in x time" eg. 18 hours. The agent is waiting until its next scheduled search and it will search when this time is reached.
- "Searching." These are shown in bold. The agent is currently searching.
- "Disabled." The agent will not perform any searches.
- "Service unavailable." The agent cannot find the database it is scheduled to search. This will happen if the database plugin has been uninstalled or if for example the Collaboration contact is offline currently.
- "No search scheduled" The agent is enabled but doesn't have a search scheduled. To correct this click the "Run now" button in the agent dialog to have it search immediately and schedule a new search.

Deliver To. This names the destination folder for the downloaded documents. This is usually your Local Documents or one of your local folders.

Note. If you close Geneious while an agent is running, it will stop in mid-search. It will resume searching when Geneious is restarted. Also, all downloaded files are stored in the destination folder and are marked "unread" until viewed for the first time.

2.6.3 Manipulating an agent

Once an agent has been set up, it can be disabled, enabled, edited, deleted and run. All these options are available from within the Agents dialog.

- *Enable or disable* an agent by clicking the check box in the Enable column.
- "Run Now" Cause the agent to search immediately
- "Cancel" If the agent is currently searching this can be clicked to stop the search.
- "Edit" Click this to change an agent's database, search criteria, destination or search interval.
- "Delete" Delete the agent permanently. Any documents retrieved by the agent will remain in your local documents.

2.7 Filtering and Similarity sorting

The "Filter" allows you to instantly identify documents in the document table matching chosen keywords. It is located in the top right hand corner of the Main Toolbar.

Type in the text you are searching for and Geneious will display all the documents that match this text and hide all other documents in the Document Table. To view all the documents in a folder, clear the Filter box of text or click the button.

The "Sort" button in the toolbar provides two actions in a popup menu. Sort by similarity is available when a single sequence document is selected in the Document Table. It will rank all other sequences by their similarity to the selected sequence. The most similar sequence is placed at the top and the least similar sequence at the bottom. This also produces an E-value column describing how similar the sequences are to the selected one. The "Remove Sort by Similarity" action will remove the E-value column and return the table to its previous sorting.

2.7.1 Filtering on-the-fly

Filtering can be used while searching for documents via public databases, filtering data as it is being downloaded. Type in the appropriate text in the Filter Box and only those documents that match both the original criteria (as specified by the search terms) and the "Filter" text will be displayed. This is an effective way of filtering within your search results.

2.8 Notes

Notes allow you to add arbitrary information to any of your local documents, and any Notes that you add can be treated as user-defined fields for use in sorting, searching and filtering your documents.

Where can I add Notes?

You can add a note to any of your local documents, including molecular sequences, phylogenetic trees and journal articles. You cannot add notes to search results from NCBI or EMBL etc until the documents are copied into one of your local folders.

The Notes View

All documents have a "Notes" tab in the document viewer panel. Click on the tab to display the Notes view, which will show you all the notes that are attached to your selected document(s). To add a note to your document, select the "Add a Note" button on the toolbar and then choose from the available note types. Selecting a note type will create an empty note of that type. To fill the note just start typing values into the fields.

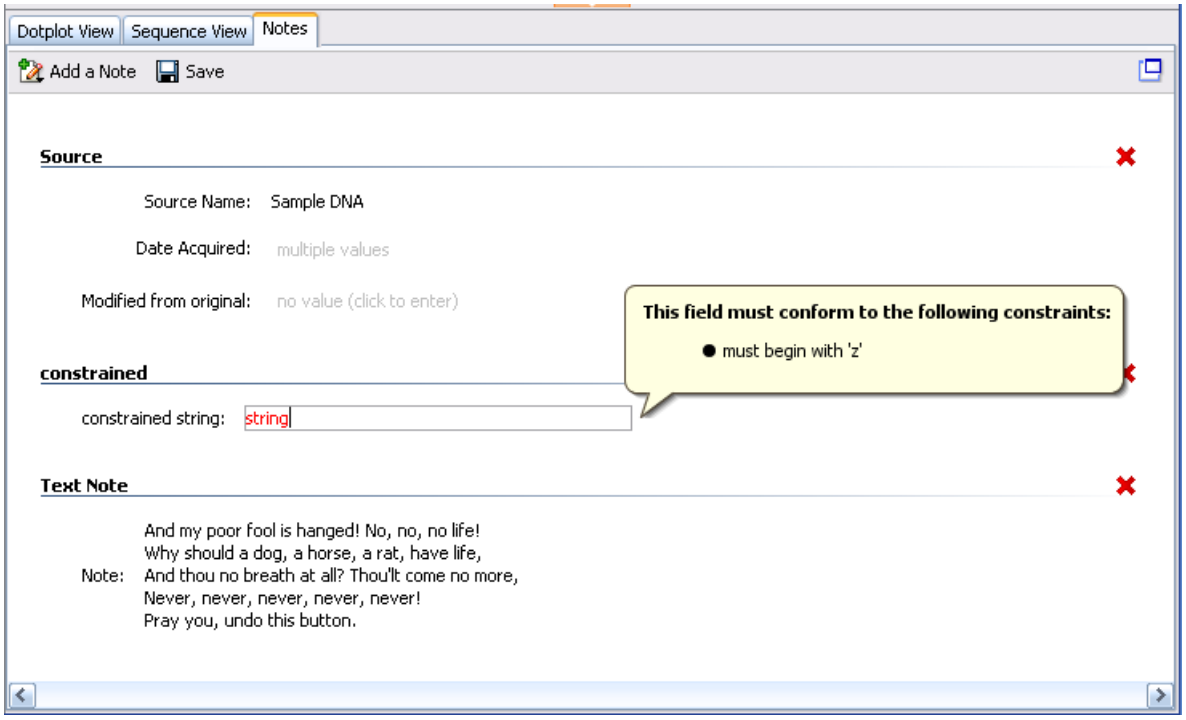

Dotplot View
Sequence View
Notes
Add a Note
Save
Source
Source Name: Sample DNA
Date Acquired: multiple values
Modified from original: no value (click to enter)
This field must conform to the following constraints:
must begin with 'z'
constrained
constrained string: string
Text Note
Note: And my poor fool is hanged! No, no, no life!
Why should a dog, a horse, a rat, have life,
And thou no breath at all? Thou'lt come no more,
Never, never, never, never, never!
Pray you, undo this button.

Figure 2.14: The Notes View

2.8.1 Editing Notes

To edit the fields of a note, simply click on the field and enter your data. Some fields may have constraints (which you can edit in the Edit Note Types dialog, *(see 2.8.2)*. If the data you have

entered does not conform to the constraints of the field, it will be displayed in red and you will be shown the field's constraints in a tooltip *(see figure 2.14)*.

***Tip: To enter a new line in a text field, press* Shift+enter *or* Ctrl+enter**

When multiple documents are selected, the Notes view displays all of the notes belonging to the selected documents. When each document has the same value for a note field, it is displayed in the viewer. If the documents have different values, or some of the selected documents do not have a note of that particular type, then the field will show that it represents multiple values. Changes made to the fields will apply to all selected documents.

2.8.2 Editing Note Types

To edit your note type, click the "Add a Note" button on the viewer toolbar and select "Edit note types...". This will bring up a window similar to that displayed in figure 2.16.

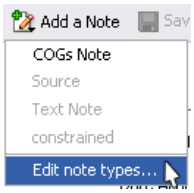


Figure 2.15: Edit Note Types

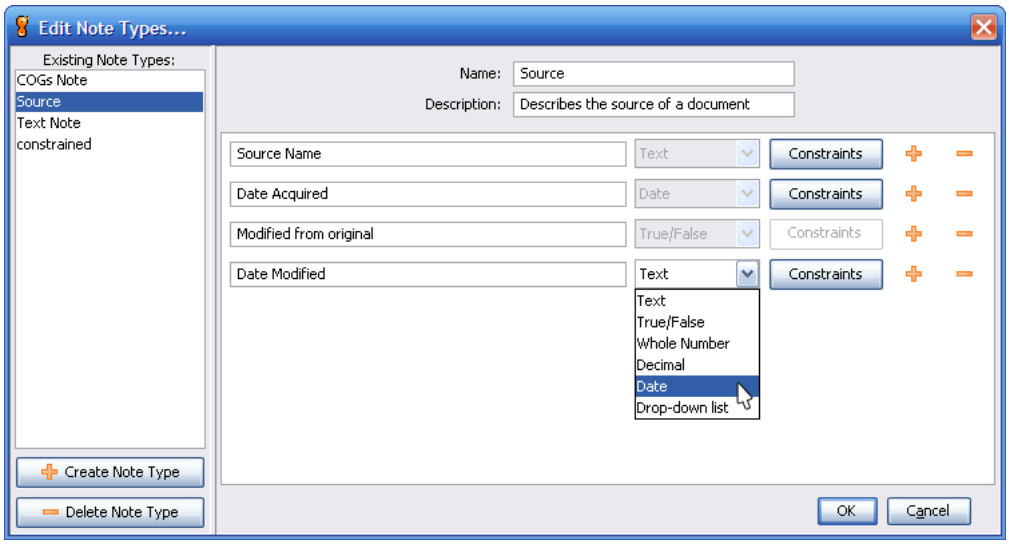


Figure 2.16: The Edit Note Types window

Creating Note Types

Geneious does not restrict you to the note types that it comes with. You can create your own note types to store any information you want.

To create a new note type, click on the Create Note Type button (+) in the left-hand panel of the Edit Note Types window. This creates a new note type, with one empty field, and displays it in the panel to the right.

Note. The "Note Type Name" and "Note Type Description" fields distinguish your Note type from other user-defined note types. They do not have any constraints. Here are some examples of Note Types.

Name	**Description**
Protein size	Size of the protein in kDa
Tree building method	Method used to build tree UPGMA/Neighbor joining

Next, you need to decide what values your Note Type will store by specifying its fields:

Field name. This defines what the field will be called. It will be displayed alongside columns such as Description and Creation Date in the Documents Table. You can have more than one Field in a single Note Type - to add or remove a field from the note type, click the + or - buttons to the right of the field.

Field type. This describes the kind of information that the column contains such as Text, Integer, and True/False. The full list of choices in Geneious is shown in figure 2.16.

Constraints. These are limiting factors on the data and are specific to each field type. For example, numbers have numerical constraints – is greater than, is less than, is greater or equal to, and is less or equal to. These can be changed to suit. The constraints for each field can be viewed by clicking the "View Constraints" button next to the field. This will show a pop-up menu with the constraints you have chosen. *(see figure 2.17)*

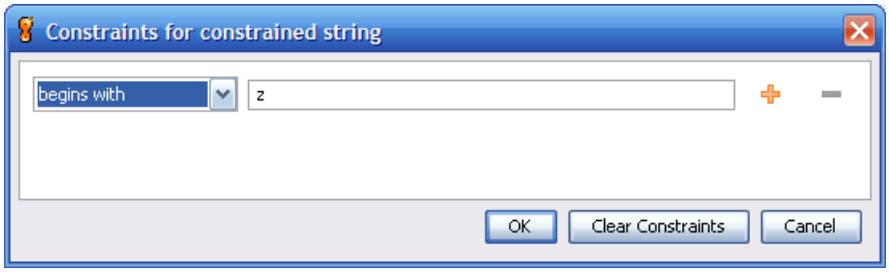


Figure 2.17: The Edit Constraints window

Using Note Types

The main purpose of Notes is to add user defined information to Geneious documents. However, Notes and Note Types can be searched for and filtered as well. Also, documents can be sorted according to the values of an added Note.

Searching - Once a Note Type is defined and a Note of that type added, it is automatically added to the standard search fields. These are listed under the "Advanced Search" options in the Document Table. From then on, you can use them to search your Local Documents. If you have more than one Field Type for a Note Type, they will both appear as searchable fields in the search criteria.

Filtering - Note values can be used to filter the documents being viewed. To do so, type a value of your Note Type into the "Filter Box" in the right hand side of the Toolbar. Only matching documents will be shown.

Sorting - The fields and values of an added Note Type will appear as columns in the Document Table. These new columns can be used to order the table. Take the example of protein size. A click on the column heading will order the documents in increasing or decreasing order according to their protein size. Clicking the column heading again arranges the documents in the opposite order. An arrow next to the heading indicates if it is in increasing ($\wedge$) or decreasing ($\vee$) order.

2.9 Preferences

You can access the preferences screen in two ways:

1. Shortcut keys: Ctrl+Shift+P (Windows/Linux), ⌘+Shift+P (Mac OS X)
2. Select the Tools Menu and click Preferences.

There are several sections in the preferences window which are presented as tabs. The most important of these are described below.

2.9.1 General

This contains connection settings, data storage details for your local documents, automatic new version checking and a "Search History".

"Check for new version of Geneious" Enable this to have Geneious check for the release of new versions everytime it is started. If a new version has been released Geneious will tell you and give you a link to download it.

"Also check for beta version of Geneious" Enable this to also have Geneious alert you when new beta versions are released. A beta version is a version that is released before the official release for the purposes of testing. It may therefore be less stable than official releases.

"Max memory available to Geneious" allows you to enter how many megabytes of your computers memory you wish to allow Geneious to use.

Search History. This clears all the previously searched words in Geneious. After this, the auto-completion drop-box will be empty.

Connection settings. These are described in the troubleshooting section of the manual.

2.9.2 Plugins and Features

The "Plugins and Features" tab (Figure 2.18) lets you manage downloadable plugins and change the features available in Geneious.

- **Available Plugins:** Lists all plugins which are currently available for download from the Geneious website which aren't already installed. Each plugin is listed with a status which can be a star (for exciting plugins), New, Pro-Only or Beta. Click the Info button to read more about the plugin or click the Install button to download the plugin and install it.
- **Installed Plugins:** Lists all plugins you currently have installed. Click the uninstall button next to a plugin to remove it.
- **Install plugin from a gplugin file:** If you have downloaded a plugin from our website or obtained one from another source (usually in .gplugin format) you can install it by clicking this button or by dragging the plugin file in to Geneious.
- **Check for plugin updates now:** Checks if there are any new versions available for the plugins you have installed.
- **Automatically check for updates to installed plugins:** If checked, Geneious will check for new versions of your installed plugins each time the program is started.
- **Tell me when new plugins are released:** Changes the way the program notifies you about new plugin releases.
- **Also check for beta releases of plugins:** Plugins are sometimes initially released as a beta for the purposes of testing before the officially release. Check this to be notified about the release of beta plugins.
- **Customize feature set:** Click this to see a list of all features in Geneious. Any number of these can be turned off by un-checking the Enabled box next to each feature. You might like to turn of the Tree Builder and Tree Viewer plugins if you don't do phylogenetics for example.

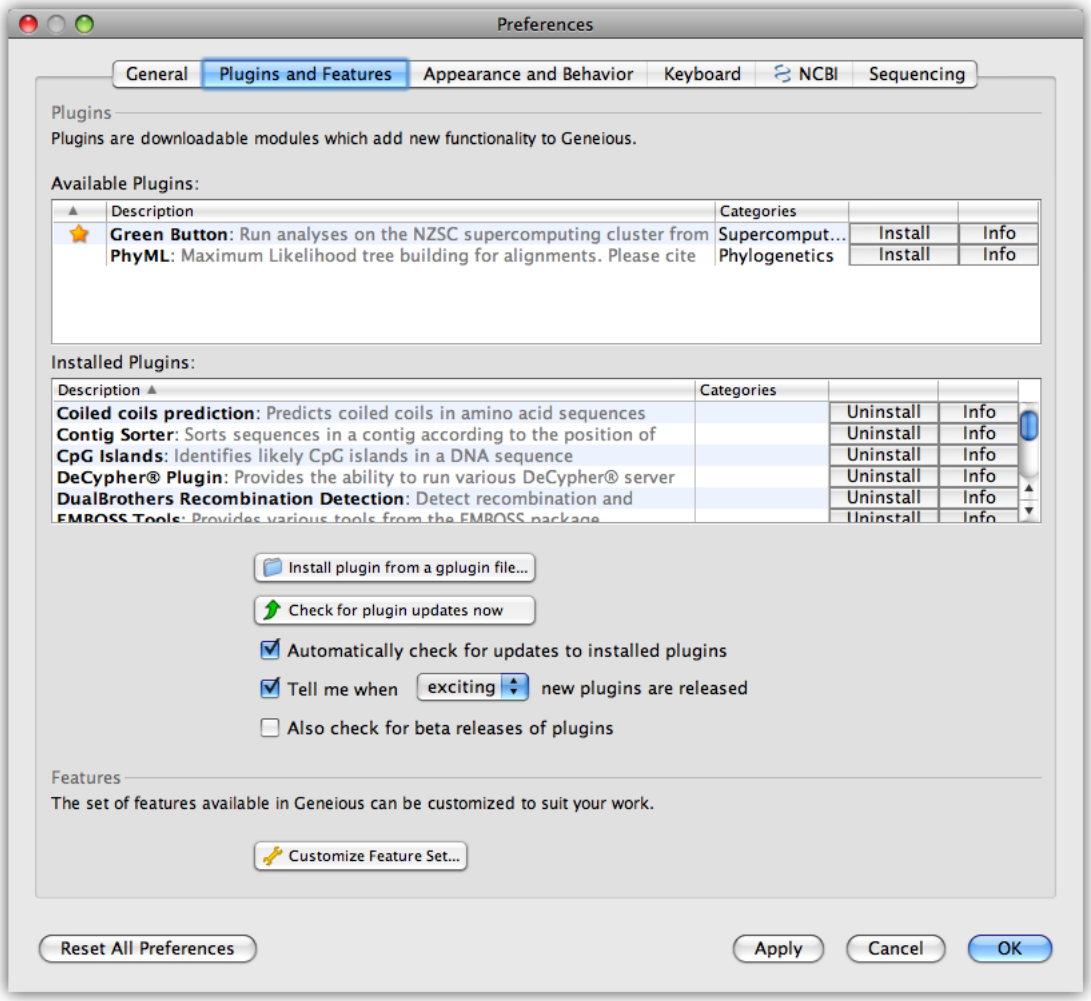


Figure 2.18: The plugins preferences in Geneious

2.9.3 Appearance and Behavior

Here you can change the way Geneious looks and the way it interacts with you.

Appearance options allow you to change the way the main toolbar and the document table look.

Behaviour options allow you to change the way newly created documents are handled. Such as whether they are selected straight away and where they should be saved to.

2.9.4 Keyboard

This section contains a list of Geneious functions and allows you to define keyboard shortcuts to them. Shortcuts that are already defined are highlighted in blue.

Setting shortcuts can help you quickly navigate through Geneious without using the mouse and also allows you to redefine shortcuts to ones you may be familiar with from other programs.

Double click on a function to bring up a window to enter your new keyboard shortcut. If you use one that is already assigned, Geneious will tell you what function currently has that shortcut.

2.9.5 Sequencing

This tab has options for the management of trace files and assemblies:

- **Confidence:** Set the threshold values of base call confidence used to determine if a base call is low, medium or high quality. This affects the binning parameters described below as well as the Confidence color scheme in the Sequence View.
- **Sequence binning options:** Specifies the requirements for individual traces to be binned as medium or high quality overall. To see the Bin for a trace, turn on the Bin column under Table Columns in the View menu.
- **Assembly binning options:** Specifies the requirements for assembly documents to be binned as medium or high quality overall. To see the Bin for an assembly, turn on the Bin column under Table Columns in the View menu.
- **Track binning history in notes:** When turned on, a note will be added to traces when they are trimmed (see the Notes view tab). This note will then updated every time the trace is re-trimmed, maintaining a history of the trimming.

2.10 Printing and Saving Images

Geneious allows you to print (or save as an image) the current display for any document viewer. This includes the sequence viewer, tree view, dotplot, and text view.

2.10.1 Printing

Choose "print" from the file menu. The following options are available

Portrait or landscape. Controls the orientation of the page.

Scale. Can be used to decrease or increase the size of everything in the view, while still printing within the same region of the page. For many types of document views, this will cause it to wrap to the following line earlier, usually requiring more pages.

Size. Controls the size the printed region on the paper. Effectively, increasing the size, reduces the margins on the page.

2.10.2 Saving Images

Choose "save to image file" from the file menu. The following options are available

Size. Controls the size of the image to be saved. Depending on the document view being saved, these may be fixed or configurable. For example, with the sequence viewer, if wrapping is on, you are able to choose the width at which the sequence is wrapped, but if wrapping is off, both the width and height will be fixed.

Format. Controls image format. Vector formats (PDF and SVG) have the advantage over raster formats (PNG and JPG) that they don't become pixelly when magnified. Vector formats are only available in the pro version.

Resolution. Only applies to raster formats (PNG and JPG) and is used to increase the number of pixels in the saved image.

Chapter 3

Document Viewers

3.1 The Sequence (and alignment) Viewer

The "Sequence view" tab in the Document Viewer panel is available for Nucleotide sequences, Protein sequences, Alignments and 3D structure documents. If an alignment is selected, this will be called "Alignment View" or "Contig View" if a contig is selected. The options available are grouped under headings: "Zoom level", "Colors", "Graphs", "Consensus & Highlighting", "Annotations", "Layout & Properties", "Complement & Translation" and "Statistics". The presence of these options varies with the kind of sequence data being viewed.

3.1.1 Zoom level

The plus and minus buttons increase and decrease the magnification of the sequence by 50%, or by 30% if the magnification is already above 50%.

zooms in to fit the selected region in the available viewing area.

zooms to 100%. The 100% zoom level allows for comfortable reading of the sequence.

zooms out so as to fit the entire sequence in the available viewing area.

Zooming can also be quickly achieved by holding down the zoom modifier key which is the Ctrl key on Windows/Linux or the Alt/Option key on Mac OS X and clicking. When the zoom key is pressed a magnifying glass mouse cursor will be displayed.

- Hold the zoom key and left click on the sequence to zoom in.
- Hold the zoom key and Shift key to zoom out.

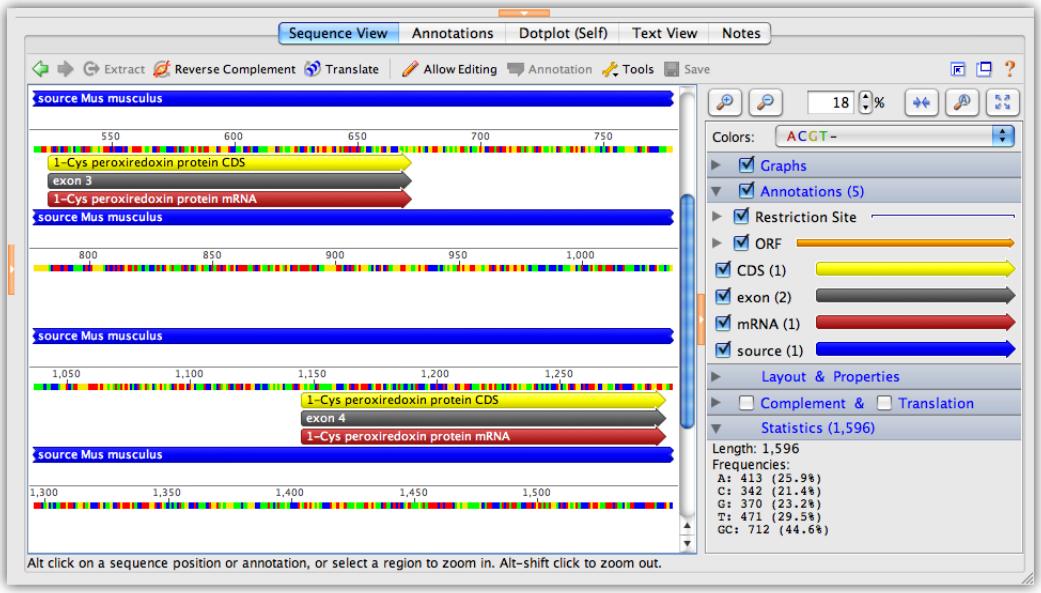


Figure 3.1: A view of an annotated nucleotide sequence in Geneious

- Hold the zoom key and turn the scroll wheel on your mouse (if you have one) to zoom in and out.
- Hold the zoom key and click on an annotation to zoom to that annotation

You can also pan in the Sequence VIew by holding Ctrl+Alt (⌘+Alt on Mac OS X) and clicking on the sequence and dragging.

3.1.2 Circular View

When a circular sequence is selected, the default view is to display the sequence as circular. The view can be rotated by using the scrollbars. Even though a sequence is circular, you can force it to display as a linear sequence using the "Linear view on circular sequence" checkbox under the "Layout" section of "Layout & Properties".

3.1.3 Colors

The colors option controls the coloring of the sequence nucleotides or amino acids. Coloring schemes differ depending on the type of sequence. For example, the "Polarity" and "Hydrophobicity" coloring schemes are available only for Protein sequences.

Similarity Color Scheme

The similarity scheme is used for quickly identifying regions of high similarity in an alignment.

In order for a column to be rendered black (100% similar) all pairs of sites in the column must have a score (according to the specified score matrix) equal to or exceeding the specified threshold.

So for example, if you have a column consisting of only K (Lysine) and R (Arginine) and are using the Blosum62 score matrix with a threshold of 1, then this column will be colored entirely black because the Blosum62 score matrix has a value of 2 for K vs R.

If you raised the threshold to 3, then this column would no longer be considered 100% similar. If the column consisted of 9 K's and 1 R, then continuing with the threshold value of 3, the 9 K's which make up 90% of the column would now be colored the dark-grey (80%-¿100%) range while the single R would remain uncolored.

If instead the column consisted of 7 K's and 3 R's (still with threshold 3) then 70% of the column is now similar so those 7 K's would be colored the lighter grey (60%-¿80%) range.

Alternatively, going back to the default threshold value of 1, and with a column consiting of 7 K's, 2 R's and 1 Y, now since the 7 K's and 2 R's have similarity exceeding the threshold whereas the Y is not that similar to K and R, the K's and R's will be colored dark grey since they make up 90% of the column.

3.1.4 Graphs

This option is visible when viewing protein sequences, chromatogram traces, multiple sequences or sequence alignments. Turn this option on by clicking the Graph checkbox and the graph(s) will be displayed below the sequence(s). The number control to the right of each graph controls the height of that graph (in pixels). A number of graphs are available.

Identity. This is available for sequence alignments. It displays the identity across all sequences for every position. Green means that the residue at the position is the same across all sequences. Yellow is for less than complete identity and red refers to very low identity for the given position (Figure 3.2).

Sequence Logo. This is available for sequence alignments. It displays a sequence logo, where the height of the logo at each site is equal to the total information at that site and the height of each symbol in the logo is proportional to its contribution to the information content. When zoomed out far enough such that he horizontal width of each site is less than one pixel, then the height is the average of the information over multiple sites. When gaps occur at at some sites, the height is scaled down further to be proportional in height to the number of non-gap residues.

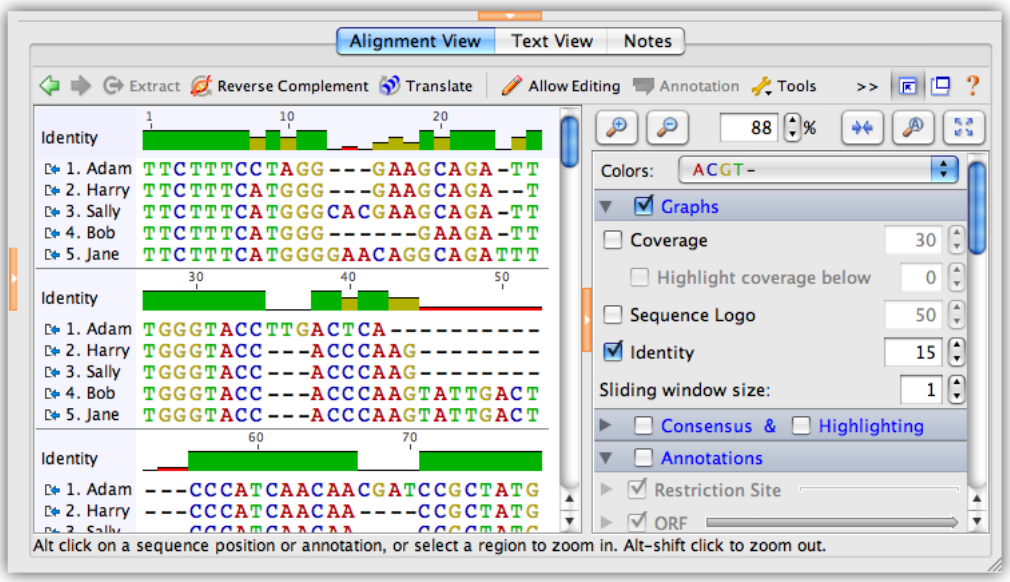


Figure 3.2: The identity graph for an alignment of nucleotide sequences

Hydrophobicity. This is available with protein sequences. It displays the Hydrophobicity of the residue at every position, or the average Hydrophobicity when there are multiple sequences.

pI. pI stands for Isoelectric point and refers to the pH at which a molecule carries no net electrical charge. The pI plot displays the pI of the protein at every position along the sequence, or the average pI when multiple sequences are being viewed.

Sliding window size. This calculates the value of the graph at each position by averaging across a number of surrounding positions. When the value is 1, no averaging is performed. When the value is 3, the value of the graph is the average of the residue value at that position and the values on either side.

Chromatogram. This is available with chromatogram traces. It displays the four traces above the sequence, where the peak as detected by the base calling program is at the middle of the base letter. When viewing more than one chromatogram or an alignment made from chromatograms, each chromatogram can be turned on or off individually using the checkbox's below. Note that since the distance between bases as inferred from the trace varies the trace may be either contracted or expanded compared with the raw data.

Quality. This is available with enabled chromatogram traces. It displays a quality measure (typically Phred quality scores) for each base as assessed by the base calling program. The quality is shown as a shaded bar graph overlaid on top of the chromatogram. Note that those scores represent an estimate of error probability and are on a logarithmic scale - the highest bar represents a one in a million (10^{-6}) probability of calling error while the middle represents a

probability of only a one in a thousand (10^{-3}).

Coverage. This is available on sequence alignments and contigs. The height of the graph at each position represents the number of sequence which have a non-gap character at that position. If the selected contig was created using Geneious and it contains sequences in both directions, then color coding is used to indicate whether each position is covered by reads in both directions. Green is used for regions with reads in both directions and yellow is used for regions with reads in one direction only.

3.1.5 Consensus & Highlighting (*Pro* only)

This option is available when viewing alignments. When checked, the viewer displays the `consensus` sequence with the aligned sequences. The consensus sequence has the same length and shows which residues are conserved (are always the same), and which residues are variable. A consensus is constructed from the most frequent residues at each site (alignment column), so that the total fraction of rows represented by the selected residues in that column reaches at least a specified threshold. IUPAC ambiguity codes (such as R for an A or G nucleotide) are counted as fractional support for each nucleotide in the ambiguity set (A and G, in this case), thus e.g. two rows with R are counted the same as one row with A and one row with G. When more than one nucleotide is necessary to reach the desired threshold, this is represented by the best-fit ambiguity symbol in the consensus; for protein sequences, this will always be an X. In the case of ties, either all or none of the involved residues will be selected. Hence, an alignment column with only As and Gs in equal number will be represented as an R in the consensus sequence regardless of the consensus threshold.

When *ignore gaps* is checked, the consensus is calculated as if each alignment column consisted only of the non-gap characters; otherwise, the gap character is treated like a normal residue, but mixing a gap with any other residue in the consensus always produces the total ambiguity symbol (N and X for nucleotides and amino acids, respectively).

When the aligned sequences contain quality information in the form of chromatograms, you can select *Highest Quality* to calculate a majority consensus that takes the relative residue quality into account.

When *Highlight disagreements* is checked, the residues in the alignment that are identical to the consensus state for that column are grayed out. This allows you to quickly locate variable sites in the alignment.

Similarly *Highlight agreements* greys out residues that are not indentical to the consensus allowing you to quickly locate conserved sites in the alignments.

Highlight ambiguities greys out non-ambiguous residues.

Highlight gaps greys out non-gap positions.

Highlight transitions/transversions greys out residues that are not transitions/transversions com-

pared to the consensus sequence. When highlighting transitions/transversions, it is recommended you turn on the ignore gaps consensus option or some residues may be wrongly highlighted due the consensus displaying N for sites that contain gaps and non-gaps.

Go to next disagreement/agreement/transition/transversion/ambiguity goes to the next highlighted feature as described in the previous section on highlighting.

Highlighting can be applied with reference to the consensus or a selected reference sequence.

3.1.6 Annotations

Some protein and nucleotide sequences come with annotations and these can be viewed within Geneious sequence viewer. In the presence of annotations, the options panel includes an "Annotations" check box (Figure 3.3). Uncheck the check box to turn off all annotations. Individual annotation types can be turned on or off by using the check boxes next to them.

Figure 3.3: The annotations options in the sequence viewer

Hide all/Show all. These buttons can be used to turn off and on all annotations on the sequence.

3.1.7 Layout & Properties

Layout & Properties has various options controlling the look of the sequence:

- *Wrap sequence*. This wraps the sequences in the viewing area.
- *Linear view on circular sequences*. This forces circular sequences to be shown linearly.
- *Spaces every 10 residues*. If you are zoomed in far enough to be able to see individual residues, then an extra white space can be seen every 10 residues when this option is selected.
- *Hide residues*. Hides the residues/bases of the sequence and just leaves the annotations visible.
- *Show Name*. Show or hide sequence and graph names inside the sequence viewer panel.
- *Show residue numbers*. This toggles the display of the residue position number above the sequence residues.
- *Show original sequence positions*. This toggles the display of the residue position numbers for the original sequence on a per sequence basis. It is only available for alignment documents and sequences that were extracted from other sequences.
- *Outline residues when zoomed out*. This adds a fine line around the sequence which can help with clarity and printing.

You can also adjust the appearance of annotations:

- *Labels*. This option changes how labels are displayed: “Inside”, “Outside”, “Inside or Outside” and “None”.
- *Overlay on bases when zoomed out*. When only a single annotation covers a region, it will be placed on top of the sequence.
- *Compress annotations*. This option reduces the vertical height of the annotations on display. This reduces the space occupied by annotations by allowing them to overlap and increases the amount of the sequence displayed on the screen.
- *Hide excessive labels*. This will reduce screen clutter by removing annotation labels which are too frequent.

Finally, you can control the size of fonts for bases, labels, names and numbering.

3.1.8 Complement & Translation

When viewing nucleotide sequences, Geneious offers complement and protein translation options.

Translations can be selected per reading frame using a range of genetic codes. They can also be created relative to selection or annotations such as CDS (Figure 3.4).

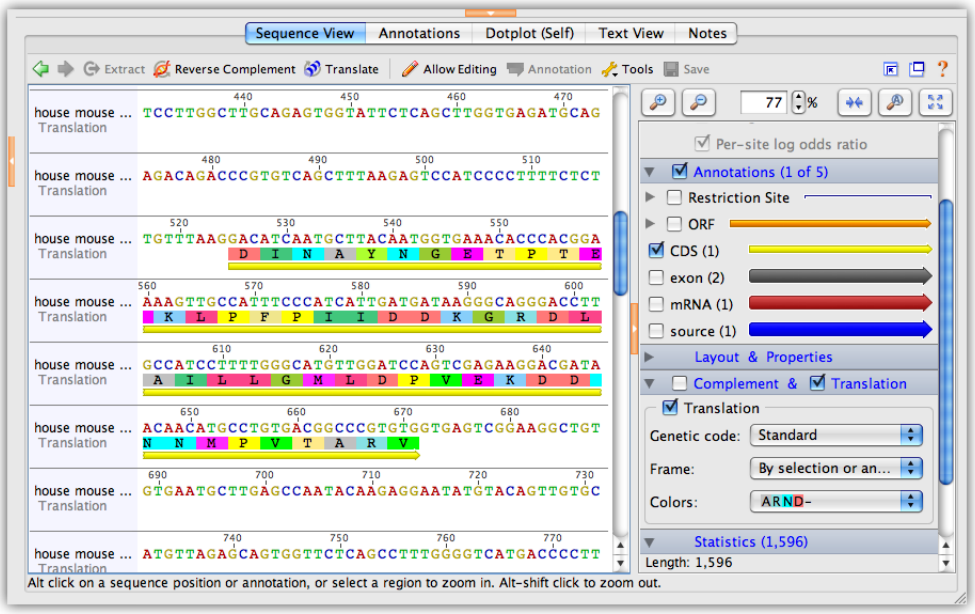


Figure 3.4: Translating a CDS

Translations can be synchronised between sequences in an alignment with reference to the individual sequences, the alignment, the consensus or a specific reference sequence.

Figure 3.5 shows an example of a DNA alignment coloured by the amino acid translation using the MacClade colour scheme.

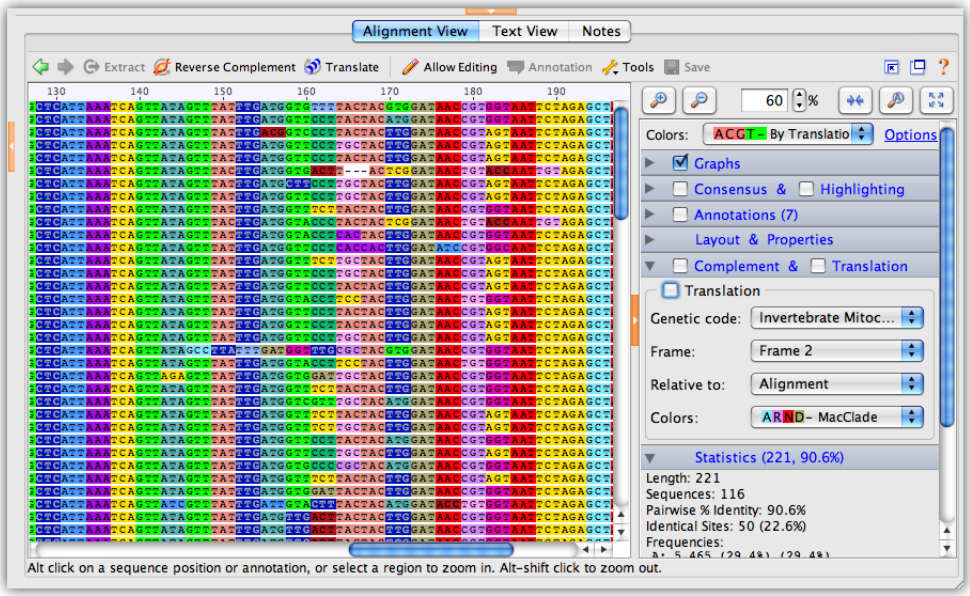


Figure 3.5: Colour an alignment by Amino Acid Translation

3.1.9 Statistics

This displays some statistics about the sequence(s) being viewed. They correspond to the sequence/alignment/assembly being viewed or the highlighted part of the sequence/alignment/assembly. The length of the sequence or part of the sequence is displayed next to the Statistics option.

Residue frequencies. This section lists the residues for both DNA and amino acid sequences, and also for alignments and assemblies. It gives the frequency of each nucleotide or amino acid over the entire length of the sequence, including gaps. If there are gaps, then a second percentage frequency is calculated ignoring gap characters. The G+C content for nucleotide sequences is shown as well for easy reference.

The following statistics are available when viewing protein sequences:

Molecular Weight. Calculates the molecular weight of the protein using the following values for the amino acids: A=71.0788 R=156.1875 N=114.1038 D=115.0886 C=103.1388 E=129.1155 Q=128.1307 G=57.0519 H=137.1411 I=113.1594 L=113.1594 K=128.1741 M=131.1926 F=147.1766 P=97.1167 S=87.0782 T=101.1051 W=186.2132 Y=163.1760 V=99.1326 U=150.0388 O=237.3018

Isoelectric Point. Calculates the isoelectric point of the protein as per this method, but using the following values for the amino acids: D=-3.9 E=-4.1 C=-8.5 Y=-10.1 H=6.5 K=10.8 R=12.5

Extinction Coefficient. Calculates the extinction coefficient of the protein as per this paper, using the following values for the amino acids and assuming all cysteines are paired in a disulfide

bridge (making cystine): C=62.5 (only counting up to an even number) W=5500 Y=1490

The following statistics are available when viewing multiple sequences:

Identical sites. When viewing alignments or assemblies this gives the percentage of columns in the alignment for which all sequences are identical.

Pairwise % Identity. When viewing alignments or assemblies this gives the average percent identity over the alignment. This is computed by looking at all pairs of bases at the same column and scoring a hit (one) when they are identical, divided by the total number of pairs.

Confidence (mean). When viewing chromatograms this gives the mean of the confidence scores for the currently selected base calls. Confidence scores are provided by the base calling program (not Geneious) and give a measure of quality (higher means a base call is more likely to be correct). An untrimmed value is also displayed if the selected region contains trims.

Expected Errors. When viewing chromatograms, this gives the approximate number of errors that are statistically expected in the currently selected region. This is calculated by converting the confidence score for each base call in to the error probability and summing across the region. This also has a value for the untrimmed selection if the region contains trims.

[Ungapped] Lengths of Sequences. Displays the mean, standard deviation, minimum and maximum of the lengths of the sequences.

Coverage of Bases. When viewing a contig assembly this gives the mean, standard deviation, mininum and maximum of the coverage of each base in the consensus sequence.

3.1.10 The sequence viewer toolbar

The top of the sequence viewer panel shows a toolbar containing several actions. Some of them operate on a part of a sequence or alignment. There are several ways to make such a selection.

- *Mouse dragging*. Click and hold down the left mouse button at the start position, and drag to the end position. By using the Ctrl (Windows/Linux) or ⌘ (Mac) keys it is possible to select multiple regions of a sequence or alignment.
- *Select from annotations* When annotations are available, click on any annotation to select the annotated residues. As with mouse dragging, multiple selections are supported.
- *Click on sequence name*. This will select the whole sequence.
- *Select all*. Use the keyboard shortcut Ctrl+A to select everything in the panel.

The available actions are,

Extract Extract the selected part of a sequence or alignment into a new document.

Reverse Complement Reverse sequence direction and replace each base by its complement. This is available only for nucleotide sequences.

Translate. Translate DNA into protein. Clicking on this choice brings up a list of genetic codes that can be used. Choose the appropriate one and click OK. This is available only for nucleotide sequences.

Allow Editing, Annotations, Tools and Save

3.1.11 Editing sequences and alignments

To edit sequence(s) or an alignment click the "Allow Editing" toolbar button. After selecting a residue or a region you can either type in the new contents or use any of the standard editing operation such as Copy (Ctrl/⌘+C), Cut (Ctrl/⌘-X), Paste (Ctrl/⌘-V) and Undo (Ctrl/⌘+Z). All operations are under the main "Edit" menu.

Selecting a region enables the "Annotations" button as well, which opens an annotation entry dialog. Enter an annotation name and select a existing type or type a new one. Click on "More Options" to enter additional properties for that annotation. Double click on an existing annotation to edit it or right-click (Ctrl+click on Mac OS X) to display a pop-up menu to delete annotations. You can also copy an annotation from one sequence to another from the pop-up menu.

When editing an alignment it is possible to select a region (which may span several sequences) and drag it to the left or right. Dragging will either move residues over existing gaps or open new gaps when necessary. Dragging a selection consisting entirely of gaps moves the gaps to the new location.

To quickly select a single residue, double-click on it. Triple clicking will select a block of residues within a single sequence. Quadruple clicking selects a block of residues in multiple sequences.

The Shift and Ctrl (Alt/Option on a Mac) keys can be combined with the keyboard arrow keys to select sequence and alignment regions. The Shift key extends the current selection and holding down the Ctrl (Alt/Option on a Mac) key while pressing the keyboard arrow is equivalent to pressing it 10 times. These can be used together. For example, in an alignment if you have a region of one sequence selected, and would like to select the same region in all sequences, then you could press Ctrl-up until you reach the first sequence, and then press Ctrl-Shift-down and few times until all sequences are selected.

Sequences can be reordered within an alignment by clicking the sequence name and dragging.

Sequences can be removed from an alignment by right-clicking (Ctrl+click on Mac OS X) on the sequence name and choosing the "remove sequence" option. Alternatively, select the entire sequence (by clicking on the sequence name) and press the delete key.

To delete a region of an alignment, select the region and press the delete or backspace key. Normally this will move residues on the right into the deleted area. By holding down the Alt key while deleting, residues on the left will be moved into the deleted area instead.

After editing is complete, click "Save" to permanently save the new contents.

3.1.12 The Pop up menu in the sequence viewer

The toolbar actions are available via a pop-up menu as well. Right-click (Ctrl+click on Mac OS X) on any sequence, partly highlighted sequence, or annotation to show the various options. The pop-up menu contains the "Copy residues" action (keyboard Ctrl+C) to copy the selected residues to the system clipboard.

3.1.13 Printing a sequence view

To print a sequence view, go to "File" → "Print" and click "OK". The view is printed without the options panel. It is recommended to turn on "Wrap sequence" and deselect "Colors" before printing. Wrapping prints the sequence as seen in the sequence viewer and the font size is chosen to fill the horizontal width of the page.

3.2 Dotplot viewer

This is a special viewer that appears when two sequences are chosen. A dotplot compares two sequences to find regions of similarity. Each axis (X and Y) on the plot represents one of the sequences being compared (Figure 3.6). For more information on dotplots, see section 4.3.

3.3 3D structure viewer

For molecular structure documents, such as PDB documents, this displays an interactive three dimensional view of the structure.

3.3.1 Structure View Manipulation

- Click and drag the mouse to rotate the structure.
- Hold the Alt or Shift key then click and drag to zoom in/out
- Hold the Ctrl key then right-click and drag to pan, or, if you are using a Mac, click and hold, press Ctrl and Alt/Option then drag to pan.

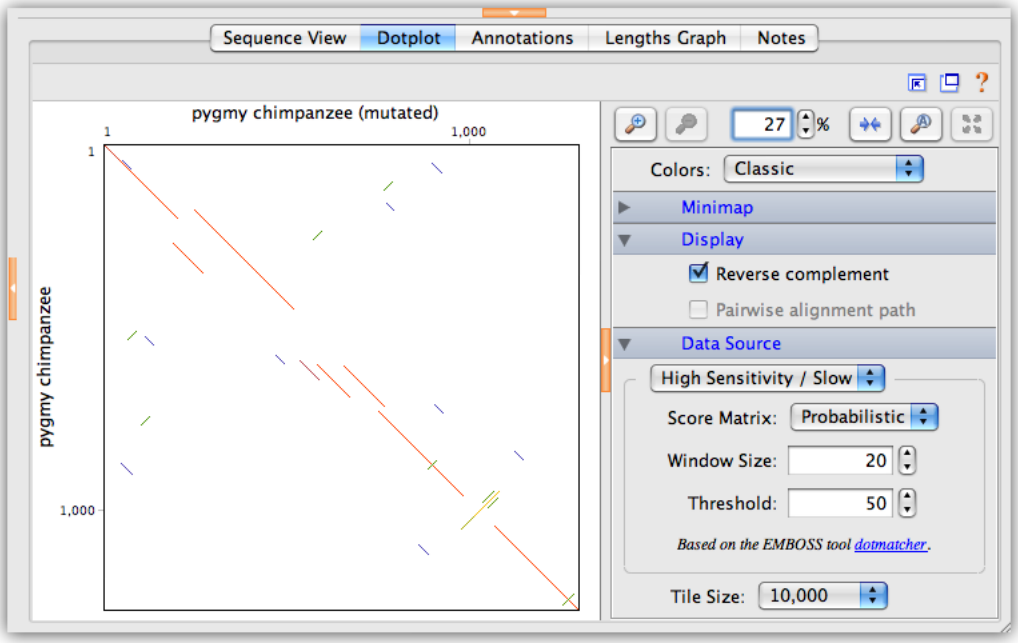


Figure 3.6: A view of dotplot of two sequences in Geneious

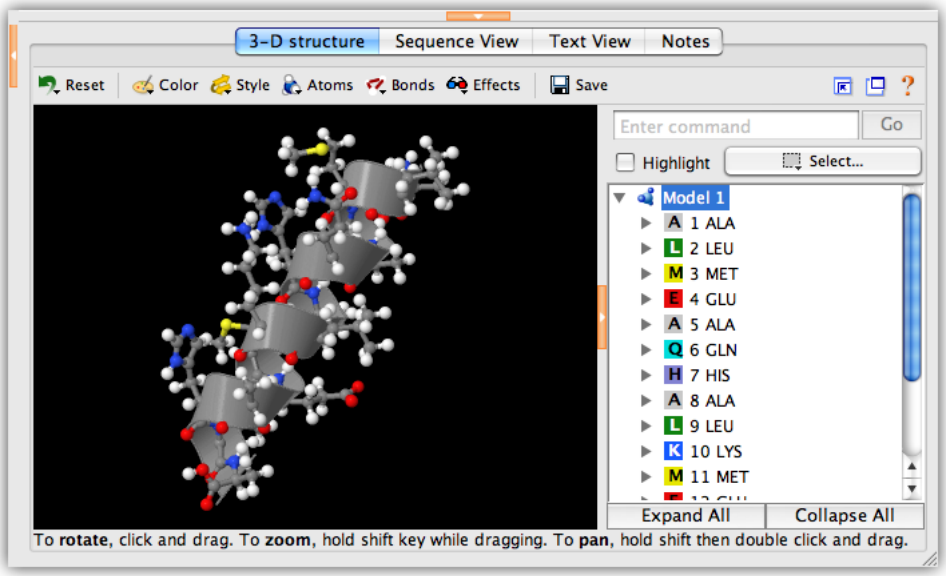


Figure 3.7: A view of a 3D protein structure in Geneious

3.3.2 Selection Controls

To the right of the structure are controls that let you control the selected part of the structure.

- If the structure you are viewing contains more than one model, the *model* combo box will you choose between them.
- The *select* button lets you select all, none or the nonselected region of the structure, as well as by element, group type or secondary structure.
- The *highlight selected* checkbox lets you select whether to highlight the selected atoms in the structure view.
- The *structure tree* shows the atoms in the structure in a tree format. Click on regions in the tree to select thoses regions. You can also Shift-click and Ctrl-click to select mutliple regions at once.
- The *command box* lets you type in arbitrary jmol scripting commands. To see some examples, select one of the pre-populated options in the box's drop-down. For a complete description of the commands you can use, see `http://www.stolaf.edu/academics/chemapps/jmol/docs`.

3.3.3 Display Menu

At the top of the viewer is the display menu. Here you can modify the appearance of the structure.

- *Reset* lets you reset the position of the structure, reset the appearance of the structure to the default, or reset the appearance of the structure to its appearance when it was last saved.
- *Color* lets you change the color scheme of the selected region of the atom.
- *Style* lets you change the style of the selected region of the molecule eg to spacefill or cartoon view.
- *Atoms* lets you hide atoms or change their size in the selected region of the molecule. You can also choose whether to show hydrogen atoms and atom symbols.
- *Bonds* lets you hide bonds or change their size in the selected region of the molecule. Covalent/ionic bonds, hydrogen bonds and disulfide bonds can be affected separately.
- *Effects* lets you toggle spin, antialiasing, stereo and slabbing effects for the whole molecule.
- *Save* saves the current appearance of the molecule.

3.4 Tree viewer

The tree viewer provides a graphical view of a phylogenetic tree (Figure 3.8). When viewing a tree a number of other view tabs may be available depending on the information at hand. The "Sequence View" tab will be visible if the tree was built from a sequence alignment using Geneious. The "Text View" shows the tree in text format (Newick).

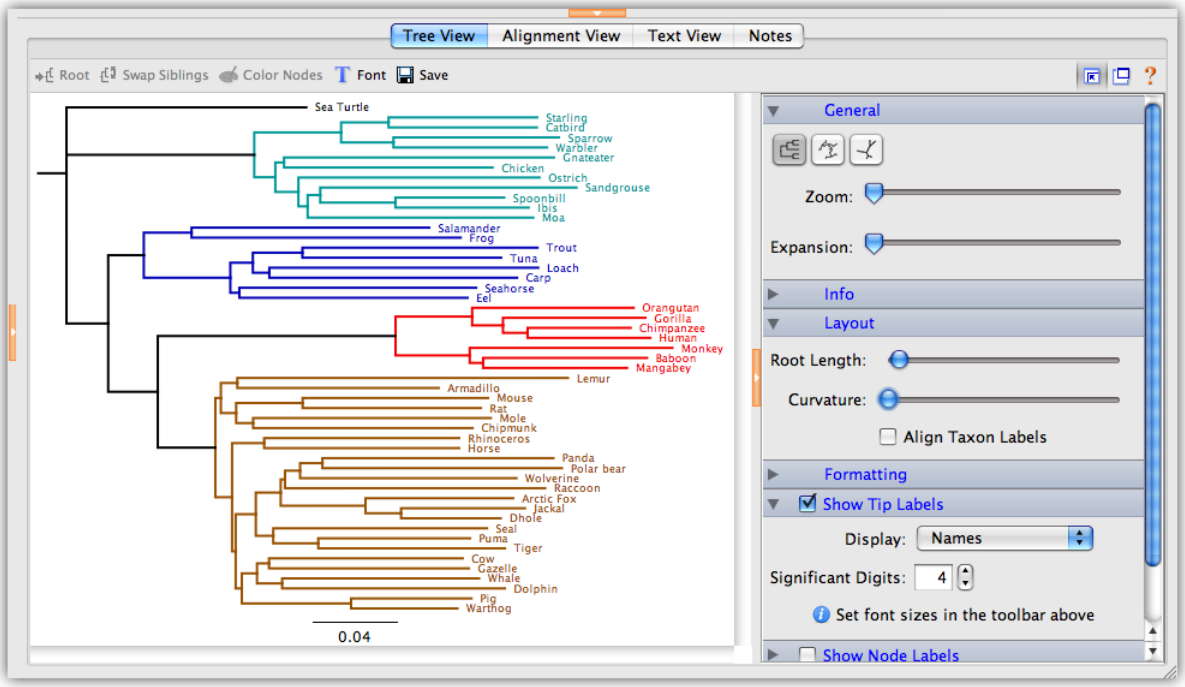


Figure 3.8: A view of a phylogenetic tree in Geneious

There are a number of options for the tree viewer.

3.4.1 Current Tree

If you are viewing a tree set, this option will be dispalyed. Select the tree you want to view from the list.

3.4.2 General

"General" has 3 buttons showing the different possible tree views: rooted, circular, and unrooted. The "Zoom" slider controls the zoom level of the tree while the "Expansion" slider

expands the tree vertically (in the rooted layout).

3.4.3 Info

For a consensus tree, the info box displays the consensus method used to build the tree. For a topology, it also shows what percentage of the original trees have the topology of the displayed tree.

3.4.4 Layout

This has different options depending on the layout that you select above:

- *Root Length* Sets the length of the visible root of the tree *(Rooted and Circular views)*
- *Curvature* Adds curvature to the tree branches *(Rooted view only)*
- *Align Taxon Labels* Aligns the tip labels to make viewing a large tree easier *(Rooted view only)*
- *Root Angle* Rotates the tree in the viewer *(Circular and Unrooted views)*
- *Angle Range* Compresses the branches into an arc *(Circular view only)*

3.4.5 Formatting

There are a range of formatting options.

Transform branches allows the branches to be equal like a cladogram, or proportional. Leaving it unselected leaves the tree in its original form.

Ordering orders branches in increasing or decreasing order of length, but within each clade or cluster.

Show root branch displays the position of the root of the tree *(has no effect in the unrooted layout)*.

Line weight can be increased or decreased to change the thickness of the lines representing the branches.

Auto subtree contract automatically contracts subtrees when there is not enough space on-screen to display them nicely.

Show selected subtree only shows only the part of the tree that is selected (or the entire tree if there is no selection).

If you are unfamiliar with tree structures, please refer to Figure 3.9 for the following options.

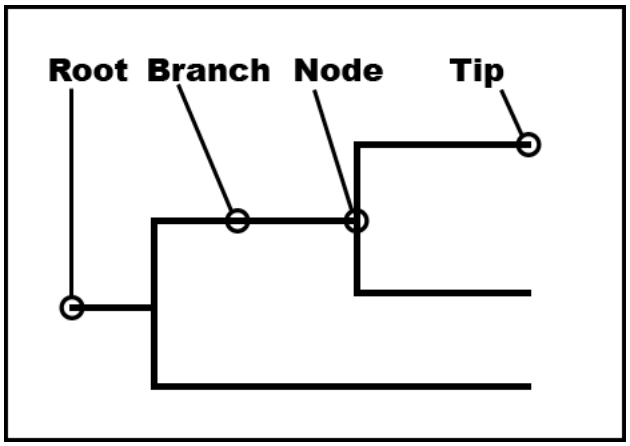


Figure 3.9: Phylogenetic tree terms

Show tip labels. This refers to labels on the tips of the branches of the tree.

Show node labels. This refers to labels on the internal nodes of the tree.

Show branch labels. This refers to the branches of the tree.

Each of the three above options has fields that you can set to customise what the labels display.

- ”Display” allows you to select what information the labels display. Branch Labels have fixed settings, but you can select what the Tip Labels display (either Taxon Names, Node Heights, Sequence Names, or a number of other options depending on the tree you are viewing). If you are viewing a consensus tree, you can also display consensus support as a percentage on node labels.

- You can use “Font” to change the size of the labels. The tree viewer will shrink the font size of some labels if they cannot all fit in the available space. ”Minimum Size” specifies the minimum size that the tree viewer is allowed to shrink the label font to.

- ”Significant Digits” sets how many digits to display if the value the node is displaying is numeric.

Show scale bar. This displays a scale bar at the bottom of the tree view to indicate the length of the branches of the tree. It has three options: “Scale range”, “font size” and “line weight”. Setting the scale range to 0.0 allows the scale bar to choose its own length, otherwise it will be the length that you specify.

3.4.6 Node Interaction

You may click on a node in the tree viewer to select the node and its clade. Double-click the node to collapse/un-collapse the clade in the view. Once you have selected a clade in the view, you may edit the tree *(see below)*

3.4.7 The Toolbar

The buttons on the toolbar along the top of the viewer allow you to edit the tree.

If you are viewing a tree made from an alignment, the "View Sequences" button allows you view the selected nodes in the sequence viewer.

The "Root" button allows you to re-root the tree on the selected node.

The "Swap Siblings" button allows you to swap the position of the sibling clades of the selected node.

3.5 The Chromatogram viewer

This viewer is hidden by default in Geneious. To turn it on, select $Tools \rightarrow Preferences$ then enable it on the "Plugins" tab.

The Chromatogram viewer provides a graphical view of a the output of a DNA sequencing machine such as Applied Biosystems 3730 DNA analyzer. The raw output of a sequencing machines is known as a `trace`, a graph showing the concentration of each nucleotide against sequence positions. The raw trace processed by a "Base Calling" software which detects peaks in the four traces and assigns the most probable base at more or less even intervals. Base calling may also assign a quality measure for each such call, typically in terms of the expected probability of making an erroneous call.

Sequence Logo. When checked, bases letters are drawn in size proportional to call quality, where larger implies better quality or smaller chance of error. Note that the scale is logarithmic: the largest base represents a one in a million (10^{-6}) or smaller probability of calling error while half of that represents a probability of only a one in a thousand (10^{-3}).

Mark calls. Draw a vertical line showing the exact location of the call made by the base calling software.

Layout. Options controlling layout and view. Those include X and Y axis scaling, size of largest base letter (when Sequence logo is on) and minimum size of base letter (to prevent bases of low quality becoming unreadable).

3.6 The PDF document viewer

To view a `.pdf` document either double click on the document in the Documents Table or click on the "View Document" button. This opens the document in an external PDF viewer such as Adobe Acrobat Reader or Preview (Mac OS X). On Linux, you can set an environmental variable named "PDFViewer" to the name of your external PDF viewer. The default viewers on Linux are `kpdf` and `evince`.

3.7 The Journal Article Viewer

This viewer provides two tabs: "Text View" and "BibTex". "Text view" displays the journal article details including the abstract. The text contains a link to the original article through Google Scholar below the title and authors (Figure 3.10). BibTex is the standard LaTeX bibliography reference and publication management data format. LaTeX is a common program used to create formatted documents including this one. The information in the BibTex screen can be exported for use in LaTeX documents.

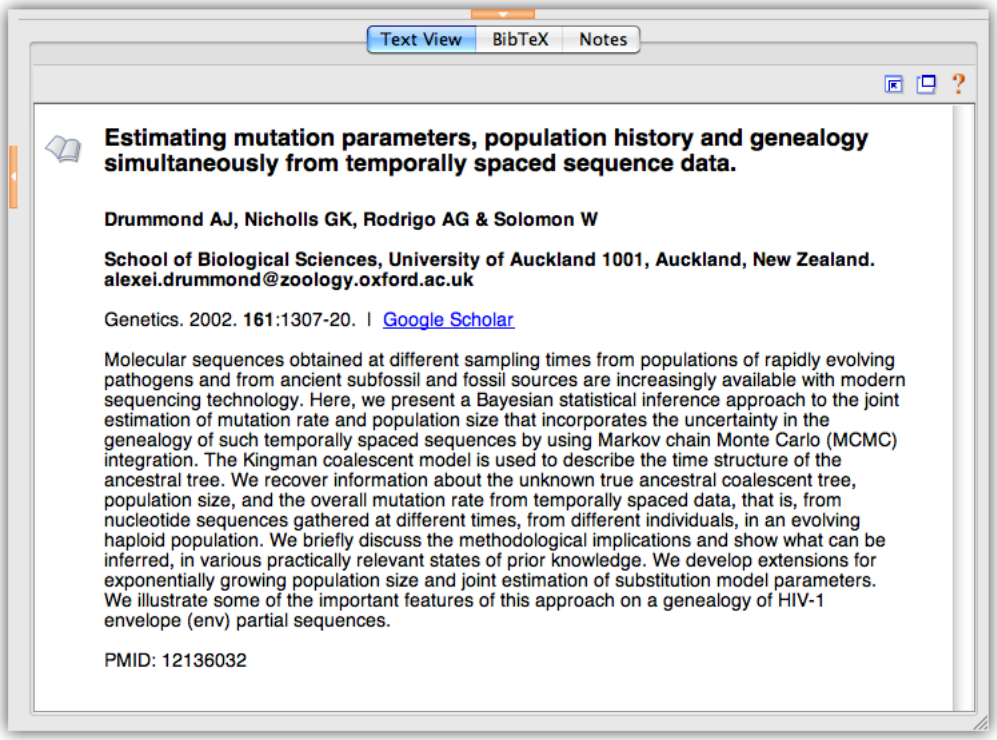

Text View
BibTeX
Notes
Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data.
Drummond AJ, Nicholls GK, Rodrigo AG & Solomon W
School of Biological Sciences, University of Auckland 1001, Auckland, New Zealand. alexei.drummond@zoology.oxford.ac.uk
Genetics. 2002. 161:1307-20. | Google Scholar
Molecular sequences obtained at different sampling times from populations of rapidly evolving pathogens and from ancient subfossil and fossil sources are increasingly available with modern sequencing technology. Here, we present a Bayesian statistical inference approach to the joint estimation of mutation rate and population size that incorporates the uncertainty in the genealogy of such temporally spaced sequences by using Markov chain Monte Carlo (MCMC) integration. The Kingman coalescent model is used to describe the time structure of the ancestral tree. We recover information about the unknown true ancestral coalescent tree, population size, and the overall mutation rate from temporally spaced data, that is, from nucleotide sequences gathered at different times, from different individuals, in an evolving haploid population. We briefly discuss the methodological implications and show what can be inferred, in various practically relevant states of prior knowledge. We develop extensions for exponentially growing population size and joint estimation of substitution model parameters. We illustrate some of the important features of this approach on a genealogy of HIV-1 envelope (env) partial sequences.
PMID: 12136032

Figure 3.10: Viewing bibliographic information in Geneious

Chapter 4

Analysing Data

4.1 Literature

Geneious allows you to search for relevant literature in NCBI's PubMed database. The results of this search are summarized in columns in the Document Table and include the PubMed ID (PMID), first and last authors, URL (if available) and the name of the Journal. When a document is selected, the abstract of the article is displayed in the Document Viewer along with a link to the full text of the document if available, and a link to Google Scholar, both below the author(s) name(s).

Note: If the full text of the article is available for download in PDF format, it can also be stored in Geneious by saving it to your hard drive and then importing it. This will allow full-text searches to be performed on the article.

As well as the abstract and links, Geneious also shows the summary of the journal article in BibTex format in a separate tab of the Document Viewer. This can be imported directly into a LaTeX document when creating a bibliography. Alternatively, a set of articles in Geneious can be directly exported to an EndNote 8.0 compatible format. This is usually done when creating a bibliography for Microsoft Word documents.

4.2 Sequence data

Basic techniques, such as dotplots and pairwise alignments, can be used to study the relationships between two sequences. However, as the number of sequences increases, methods for determining the evolutionary relationships between them become more complicated.

When analyzing more than two sequences, there are some common steps to determine the ancestral relationships between them. The following sections outline the basic tools for prelim-

inary sequence analysis: dotplots, sequence alignment and phylogenetic tree building.

4.3 Dotplots

A dotplot compares two sequences against each other and helps identify similar regions [14]. Using this tool, it can be determined whether a similarity between the two sequences is global (present from start to end) or local (present in patches).

The Geneious dotplot offers two different comparison engines based on the EMBOSS `dottup` and `dotmatcher` programs. The former is much faster but less sensitive than the latter. More information on these programs can be found by going to `http://emboss.sourceforge.net`.

When viewing a pairwise alignment you can activate the path which shows where the pairwise alignment runs through the dotplot. Also, for nucleotide comparisons you can show the reverse complement.

4.3.1 Viewing Dotplots

To view a dotplot in Geneious, select two nucleotide or protein sequences in the Document Table and select Dotplot Viewer in the Document Viewer Panel (Figure 3.6). The Dotplot Viewer allows you to zoom in and out, and to customize sensitivity of the comparison.

If a single nucleotide or protein sequence is selected then the dotplot is also available. In this case it shows a comparison of the sequence to itself.

The dotplot comparison of two sequences is drawn from top-left to bottom-right in and offers a selection of different color schemes. There is also a minimap available which aids navigation of large dotplots by showing the overall comparison and a box indicating where the dotplot window sits.

4.3.2 Interpreting a Dotplot

- Each axis of the plot represents a sequence.
- A long, largely continuous, diagonal indicates that the sequences are related along their entire length.
- Sequences with some limited regions of similarity will display short stretches of diagonal lines.
- Diagonals on either side of the main diagonal indicate repeat regions caused by duplication.

- A random scattering of dots reflects a lack of significant similarity. These dots are caused by short sub-sequences that match by chance alone.

For more information on dotplots, refer to the paper by Maizel & Lenk [14].

4.4 Sequence Alignments

Over evolutionary time, related DNA or amino acid sequences diverge through the accumulation of mutation events such as nucleotide or amino acid substitutions, insertions and deletions.

A *sequence alignment* is an attempt to determine regions of homology in a set of sequences. It consists of a table with one sequence per row, and with each column containing homologous residues from the different sequences, i.e. residues that are thought to have evolved from a common ancestral nucleotide/amino acid. If it is thought that the ancestral nucleotide/amino acid got lost on the evoluationary path to one descendant sequence, this sequence will show a special gap character "–" in that alignment column.

4.4.1 Pairwise sequence alignments

There are two types of pairwise alignments: *local* and *global* alignments.

A Local Alignment. A local alignment is an alignment of two sub-regions of a pair of sequences [21]. This type of alignment is appropriate when aligning two segments of genomic DNA that may have local regions of similarity embedded in a background of a non-homologous sequence.

A Global Alignment. A global alignment is a sequence alignment over the entire length of two or more nucleic acid or protein sequences. In a global alignment, the sequences are assumed to be homologous along their entire length [16].

Scoring systems in pairwise alignments

In order to align a pair of sequences, a scoring system is required to score matches and mismatches. The scoring system can be as simple as "+1" for a match and "-1" for a mismatch between the pair of sequences at any given site of comparison. However substitutions, insertions and deletions occur at different rates over evolutionary time. This variation in rates is the result of a large number of factors, including the mutation process, genetic drift and natural selection. For protein sequences, the relative rates of different substitutions can be empirically determined by comparing a large number of related sequences. These empirical measurements can then form the basis of a scoring system for aligning subsequent sequences. Many scoring

systems have been developed in this way. These matrices incorporate the evolutionary preferences for certain substitutions over other kinds of substitutions in the form of log-odd scores. Popular matrices used for protein alignments are BLOSUM [10] and PAM [2] matrices.

Note: The BLOSUM matrix is a substitution matrix. The number of a BLOSUM matrix indicates the threshold (%) similarity between the sequences originally used to create the matrix. BLOSUM matrices with higher numbers are more suitable for aligning closely related sequences.

When aligning protein sequences in Geneious, a number of BLOSUM and PAM matrices are available.

Algorithms for pairwise alignments

Once a scoring system has been chosen, we need an algorithm to find the optimal alignment of two sequences. This is done by inserting gaps in order to maximize the alignment score. If the sequences are related along their entire sequence, a global alignment is appropriate. However, if the relatedness of the sequences is unknown or they are expected to share only small regions of similarity, (such as a common domain) then a local alignment is more appropriate.

An efficient algorithm for global alignment was described by Needleman and Wunsch [16], and their algorithms was later extended by Gotoh to model gaps more accurately [6]. For local alignments, the Smith-Waterman algorithm [21] is the most commonly used. See the references provided for further information on these algorithms.

Pairwise alignment in Geneious

A dotplot is a comparison of two sequences. A pairwise alignment is another such comparison with the aim of identifying which regions of two sequences are related by common ancestry and which regions of the sequences have been subjected to insertions, deletions, and substitutions.

The options available for the alignment cost matrix will depend on the kind of sequence.

- Protein sequences have a choice of PAM [2] and BLOSUM [10] matrices.
- Nucleotide sequences have choices for a pair of match/mismatch costs. Some scores distinguish between two types of mismatches: transition and transversion. Transitions ($A \leftrightarrow G$, $C \leftrightarrow T$) generally occur more frequently than transversions. Differences in the ratio of transversions and transversions result in various models of substitution. When applicable, Geneious indicates the target sequence similarity for the alignment scores, i.e. the amount of similarity between the sequences for which those scores are optimal.
- Both protein and nucleotide pairwise alignments have choices for gap open / gap extension penalties/costs. Unlike many alignment programs these values are not restricted to

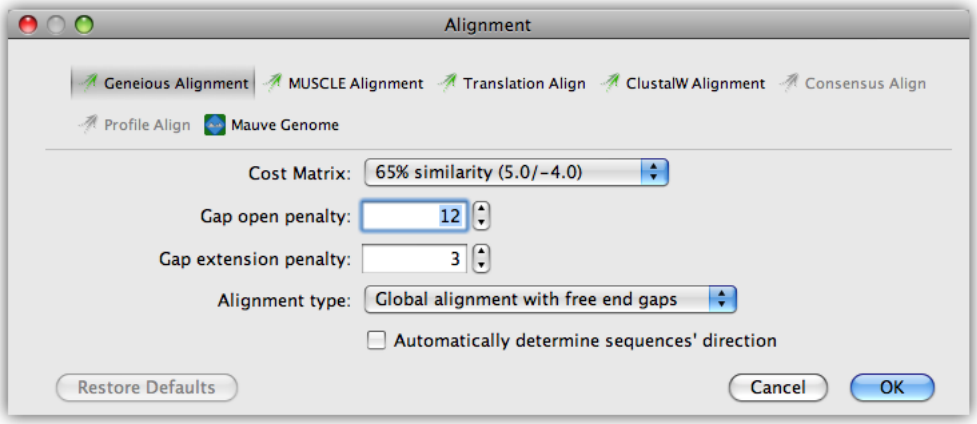


Figure 4.1: Options for nucleotide pairwise alignment

integers in Geneious.

The score of a pairwise alignment is *matchCount*matchCost+mismatchCount*mismatchCost*.

For each gap of length n, a score of *gapOpenPenalty+(n-1)*gapExtensionPenalty* is subtracted from this.

Where

- gapOpenPenalty = The "gap open penalty" setting in Geneious.
- gapExtensionPenalty = The "gap extension penalty" setting in Geneious.
- matchCost = The first number in the Geneious cost matrix.
- mismatchCost = The second number in the Geneious cost matrix.
- matchCount = The number of matching residues in the alignment.
- mismatchCount = The number of mismatched residues in the alignment.

When doing a *Global alignment with free end gaps*, gaps at either end of the alignment are not penalized when determining the optimal alignment. This is especially useful if you are aligning sequence fragments that overlap slightly in their starting and ending positions (e.g. when using two slightly different primer pairs to extract related sequence fragments from different samples). You can also do a *Local Alignment* if you want to allow free end overlaps, rather than just free end gaps in one alignment.

If you are aligning nucleotide sequences, you will also have the option of doing your alignment by translation and back. To view the options for translation alignment, click the *More Options*

button that the bottom of the alignment dialog. The translation alignment options will appear. Here you can set the genetic code and translation frame for the translation as well as the cost matrix, gap open penalty and gap extension penalty for the alignment. If you want to set the alignment type (global or local) or choose to automatically determine the sequences' direction, do it in the main section of the dialog.

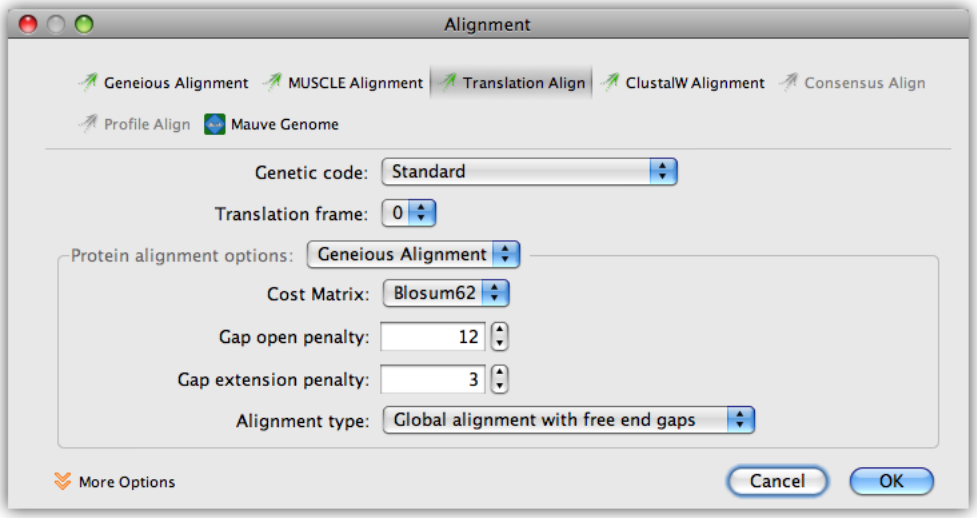


Figure 4.2: Options for nucleotide translation alignment

4.4.2 Multiple sequence alignments

A multiple sequence alignment is a comparison of multiple related DNA or amino acid sequences. A multiple sequence alignment can be used for many purposes including inferring the presence of ancestral relationships between the sequences. It should be noted that protein sequences that are structurally very similar can be evolutionarily distant. This is referred to as distant homology. While handling protein sequences, it is important to be able to tell what a multiple sequence alignment means – both structurally and evolutionarily. It is not always possible to clearly identify structurally or evolutionarily homologous positions and create a single "correct" multiple sequence alignment [3].

Multiple sequence alignments can be done by hand but this requires expert knowledge of molecular sequence evolution and experience in the field. Hence the need for automatic multiple sequence alignments based on objective criteria. One way to score such an alignment would be to use a probabilistic model of sequence evolution and select the alignment that is most probable given the model of evolution. While this is an attractive option there are no efficient algorithms for doing this currently available. However a number of useful heuristic algorithms for multiple sequence alignment do exist.

Progressive pairwise alignment methods

The most popular and time-efficient method of multiple sequence alignment is progressive pairwise alignment. The idea is very simple. At each step, a pairwise alignment is performed. In the first step, two sequences are selected and aligned. The pairwise alignment is added to the mix and the two sequences are removed. In subsequent steps, one of three things can happen:

- Another pair of sequences is aligned
- A sequence is aligned with one of the intermediate alignments
- A pair of intermediate alignments is aligned

This process is repeated until a single alignment containing all of the sequences remains. Feng & Doolittle were the first to describe progressive pairwise alignment [5]. Their algorithm used a guide tree to choose which pair of sequences/alignments to align at each step. Many variations of the progressive pairwise alignment algorithm exist, including the one used in the popular alignment software ClustalX [23].

Multiple sequence alignment in Geneious

Multiple sequence alignment in Geneious is done using progressive pairwise alignment. The neighbor-joining method of tree building is used to create the guide tree.

As progressive pairwise alignment proceeds via a series of pairwise alignments this function in Geneious has all the standard pairwise alignment options. In addition, Geneious also has the option of refining the multiple sequence alignment once it is done. “Refining” an alignment involves removing sequences from the alignment one at a time, and then realigning the removed sequence to a ”profile” of the remaining sequences. The number of times each sequence is realigned is determined by the “refinement iterations” option in the multiple alignment window. The resulting alignment is placed in the folder containing the sequences aligned.

A profile is a matrix of numbers representing the proportion of symbols (nucleotide or amino acid) at each position in an alignment. This can then be pairwise aligned to another sequence or alignment profile. When pairwise aligning profiles, mismatch costs are weighted proportional to the fraction of mismatching bases and gap introduction and gap extension costs are proportionally reduced at sites where the other profile contains some gaps.

In some cases building a guide tree can take a long time since it requires making a pairwise alignment between each pair of sequences. The “build guide tree via alignment” option may speed this part by taking a different route. First make a progressive multiple alignment using a random ordering, and use that alignment to build the guide tree. Notice that while this typically speeds up the process that may not be the case when the sequences are very distant genetically.

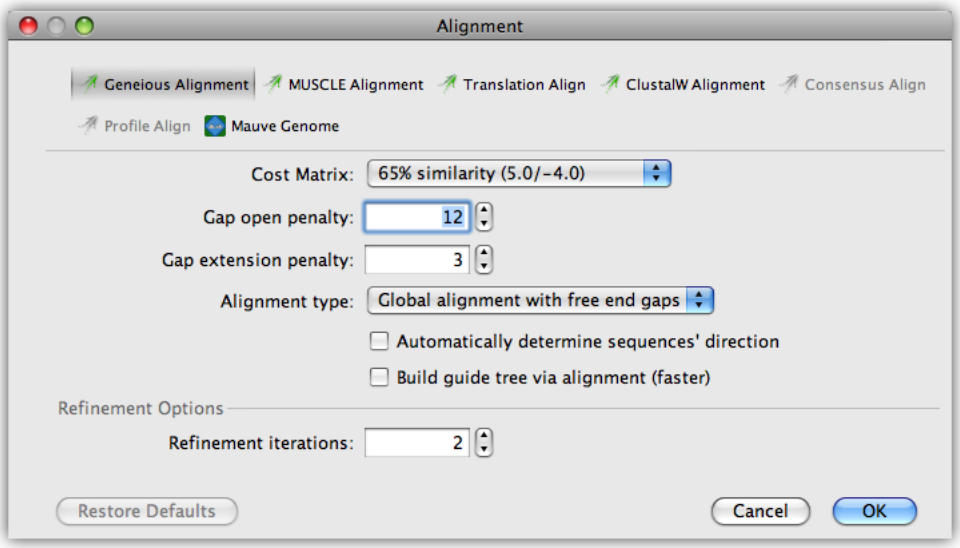


Figure 4.3: The multiple alignment window

You can also do a multiple alignment via translation and back, as with pairwise alignment.

4.4.3 Sequence alignment using ClustalW (*Pro* only)

ClustalW is a widely used program for performing sequence alignment [24, 23]. If you have ClustalW installed your computer, *Geneious pro* allows you to run ClustalW directly from inside the program without having to export or import your sequences.

If you do not have ClustalW or are unsure if you do, you should attempt to perform a ClustalW alignment without specifying a location. Geneious will then present you with options including details on how to download ClustalW, and will offer to automatically search for ClustalW on your hard drive.

To perform an alignment using ClustalW, select the sequences or alignment you wish to align and select the "Alignment" button from the Toolbar. At the top of the alignment options window ,there are buttons allowing you to select the type of alignment you wish to do. Choose "ClustalW" here, and the options available for a ClustalW alignment will be displayed.

The options are:

- *ClustalW Location:* This should be set to the location of the ClustalW program on your computer. Enter the path to it in the text field or click the "Browse" button to browse for the location. If the location is invalid and you attempt to perform an alignment Geneious

will tell you and offer the options detailed above for getting or finding ClustalW.

- *Cost Matrix:* Use this to select the desired cost matrix for the alignment. The available options here will change according to the type of the sequences you wish to align. You can also click the "Custom File" button to use a cost matrix that you have on your computer (the format of these is the same as for the program BLAST).
- *Gap open cost* and *Gap extend cost:* Enter the desired gap costs for the alignment.
- *Free end gaps:* Select this option to avoid penalizing gaps at either end of the alignment. See details in the Pairwise Alignment section above.
- *Preserve original sequence order:* Select this option to have the order of the sequences in the table preserved so that the alignment contains the sequences in the same order.
- *Additional options:* Any additional parameters accepted by the ClustalW command line program can be entered here. Refer to the ClustalW manual for a description of the available parameters.

You can also do a clustal alignment via translation and back, as with pairwise alignment.

After entering the desired options click "OK" and ClustalW will be called to align the selected sequences or alignment. Once complete, a new alignment document will be generated with the result as detailed previously.

4.4.4 Sequence alignment using MUSCLE (*Pro* only)

MUSCLE is public domain multiple alignment software for protein and nucleotide sequences. MUSCLE stands for multiple sequence comparison by log-expectation. See http://www.drive5.com/muscle/.

To perform an alignment using MUSCLE, select the sequences or alignment you wish to align and select the "Alignment" button from the Toolbar. At the top of the alignment options window ,there are buttons allowing you to select the type of alignment you wish to do. Choose "MUSCLE" here, and the options available for a MUSCLE alignment will be displayed.

For more information on muscle and its options, please refer to the original documentation for the program: http://www.drive5.com/muscle/muscle.html.

4.4.5 Combining alignments and adding sequences to alignments

Two alignment methods are available in Geneious which allow you to align two alignments together (and create a single alignment) and align a new sequence in to an existing alignment. These are "Profile Alignment" and "Consensus Alignment".

To perform either of these, select the sequences or alignment you wish to align and select the "Alignment" button from the Toolbar. At the top of the alignment options window ,there are buttons allowing you to select the type of alignment you wish to do. Choose "Profile Align" or "Consensus Align" here, and the options available for your chosen alignment will be displayed.

Profile Alignment

is performed by creating a "profile" of each alignment/sequence and then aligning those using the Needleman-Wunsch [16] global pairwise alignment algorithm. Profiles are described in the multiple alignment section (4.4.2.)

Consensus Alignment

operates in a similar way to profile alignment except it generates a consensus sequence for each alignment instead of a profile. Consensus alignment allows you to choose which alignment algorithm to use for aligning the consensus sequences. All of the pairwise and multiple alignment algorithms are available.

The consensus sequence used for each alignment is a 100% consensus with gaps ignored.

4.5 Building Phylogenetic trees

Geneious provides some basic phylogenetic tree reconstruction algorithms for a preliminary investigation of relationships between newly acquired sequences. For more sophisticated methods of phylogenetic reconstruction such as Maximum Likelihood and Bayesian MCMC we recommend specialist software such as MrBayes [19] and PhyML [7] which are available as a plugins to Geneious. These can be downloaded from the plugins page on our website.

Geneious implements the Neighbor-joining [20] and UPGMA [15] methods of tree reconstruction.

4.5.1 Phylogenetic tree representation

A phylogenetic tree describes the evolutionary relationships amongst a set of sequences. They have a few commonly associated terms that are depicted in Figure 3.9 and are described below.

Branch length. A measure of the amount of divergence between two nodes in the tree. Branch lengths are usually expressed in units of substitutions per site of the sequence alignment.

Nodes or internal nodes of a tree represent the inferred common ancestors of the sequences that are grouped under them.

Tips or leaves of a tree represent the sequences used to construct the tree.

Taxonomic units. These can be species, genes or individuals associated with the tips of the tree.

A phylogenetic tree can be rooted or unrooted. A rooted tree consists of a root, or the common ancestor for all the taxonomic units of the tree. An unrooted tree is one that does not show the position of the root. An unrooted tree can be rooted by adding an outgroup (a species that is distantly related to all the taxonomic units in the tree). A common format for representing phylogenetic trees is the Newick format [13].

4.5.2 Neighbor-joining

In this method, neighbors are defined as a pair of leaves with one node connecting them. The principle of this method is to find pairs of leaves that minimize the total branch length at each stage of clustering, starting with a star-like tree. The branch lengths and an unrooted tree topology can quickly be obtained by using this method without assuming a molecular clock [20].

4.5.3 UPGMA

This clustering method is based on the assumption of a molecular clock [15]. It is appropriate only for a quick and dirty analysis when a rooted tree is needed and the rate of evolution is does not vary much across the branches of the tree.

4.5.4 Distance models or molecular evolution models for DNA sequences

The evolutionary distance between two DNA sequences can be determined under the assumption of a particular model of nucleotide substitution. The parameters of the substitution model define a rate matrix that can be used to calculate the probability of evolving from one base to another in a given period of time. This section briefly discusses some of the substitution models available in Geneious. Most models are variations of two sets of parameters – the *equilibrium frequencies* and *relative substitution rates*.

Equilibrium frequencies refer to the background probability of each of the four bases A, C, G, T in the DNA sequences. This is represented as a vector of four probabilities $\pi_A, \pi_C, \pi_G, \pi_T$ that sum to 1.

Relative substitution rates define the rate at which each of the transitions ($A \leftrightarrow G$, $C \leftrightarrow T$) and transversions ($A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$, $G \leftrightarrow T$) occur in an evolving sequence. It is represented

as a 4x4 matrix with rates for substitutions from every base to every other base.

Additionally, *gaps* are not penalized when using the Geneious Tree Builder. Comparisons involving any gaps are ignored when calculating the distance matrix.

Jukes Cantor

This is the simplest substitution model [11]. It assumes that all bases have the same equilibrium base frequency, i.e. each nucleotide base occurs with a frequency of 25% in DNA sequences and each amino acid occurs with a frequency of 5% in protein sequences. This model also assumes that all nucleotide substitutions occur at equal rates and all amino acid replacements occur at equal rates.

HKY

The HKY model [9] assumes every base has a different equilibrium base frequency, and also assumes that transitions evolve at a different rate to the transversions.

Tamura-Nei

This model also assumes different equilibrium base frequencies. In addition to distinguishing between transitions and transversions, it also allows the two types of transitions ($A \leftrightarrow G$ and $C \leftrightarrow T$) to have different rates [22].

4.5.5 Resampling – Bootstrapping and jackknifing

Resampling is a statistical technique where a procedure (such as phylogenetic tree building) is repeated on a series of data sets generated by sampling from one original data set. The results of analyzing the sampled data sets are then combined to generate summary information about the original data set.

In the context of tree building, resampling involves generating a series of sequence alignments by sampling columns from the original sequence alignment. Each of these alignments (known as pseudoreplicates) is then used to build an individual phylogenetic tree. A consensus tree can then be constructed by combining information from the set of generated trees or the topologies that occur can be sorted by their frequency (see below). [4].

Bootstrapping is the statistical method of resampling with replacement. To apply bootstrapping in the context of tree building, each pseudo-replicate is constructed by randomly sampling columns of the original alignment with replacement until an alignment of the same size is obtained [4].

Jackknifing is a statistical method of numerical resampling based on deleting a portion of the original observations for each pseudo-replicate. A 50% jackknife randomly deletes half of the columns from the alignment to create each pseudo-replicate.

4.5.6 Consensus trees

A consensus tree provides an estimate for the level of support for each clade in the final tree. It is built by combining clades which occurred in at least a certain percentage of the resampled trees. This percentage is called the consensus support threshold. A 100% support threshold results in a *"Strict consensus tree"* which is a tree where the included clades are those that are present in all the trees of the original set. A 50% threshold results in a *"Majority rule consensus tree"* that includes only those clades that are present in the majority of the trees in the original set. A threshold less that 50% gives rise to a *"Greedy consensus tree"*. In constructing a *"Greedy consensus tree"* clades are first ordered according to the number of times they appear (i.e. the amount of support they have), then the consensus tree is constructed progressively to include all those clades whose support is above the threshold **and** that are compatible with the tree constructed so far.

Note: The above definitions apply to rooted trees. The same principles can be applied to unrooted trees by replacing "clades" with "splits". Each branch (edge) in an unrooted tree corresponds to a different split of the taxa that label the leaves of this tree.

4.5.7 Sort topologies

This will produce one or more trees summarizing the results of resampling. The frequency of each topology in the set of original trees is calculated and the topologies are sorted by their frequency. A number of these topologies, based on the topology threshold, will be output as summary trees. The summary trees have branch lengths that are the average of the lengths of the same branch from trees with the same topology.

The topology threshold determines what percentage of the original tree topologies must be represented by the summarizing topologies. The most common topology will always be output as the first summary tree. If the frequency (%) of this does not meet the threshold then the next most frequent topology will be added, and so on until the total frequency of the topologies reaches the threshold value.

A topology threshold of 0 will result in only the most common topology being output, a threshold of 100 will result in all topologies being output.

4.5.8 Tree building in Geneious

Geneious can build a phylogenetic tree for a set of sequences using pairwise genetic distances. To build a tree, select an alignment or a set of related sequences (all DNA or all protein) in the Document table and click the "Tree" icon or choose this option from the Tools menu.

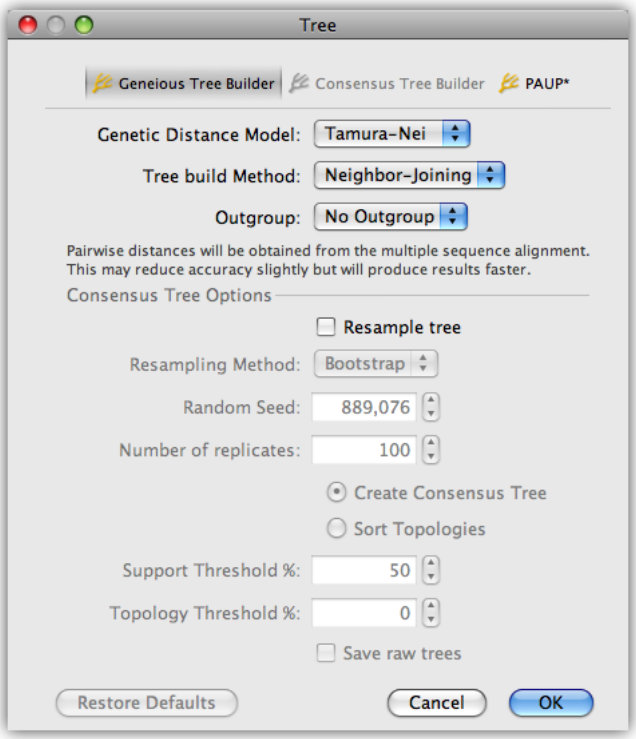


Figure 4.4: Tree building options in Geneious

Tree building from an alignment

If you are building a tree from an alignment, the following options are seen in the tree window.

If you select a tree document (which contains an alignment) then the alignment will simply be extracted from the tree and used in the tree building process.

- *Genetic distance model*. This lets the user choose the kind of substitution model used to estimate branch lengths. If you are building a tree from DNA sequences you have the choices "Jukes Cantor", "HKY" and "Tamura Nei". If you are building a tree from amino acid sequences you only have the option of "Jukes Cantor" distance correction.

- *Tree building method*. There are two methods under this option – Neighbor joining [20] and UPGMA [15].
- *Create consensus via resampling*. Check this box to build a consensus tree using resampling of sequence alignment data.
- *Resample tree* Check this to perform resampling.
- *Resampling method*. Either bootstrapping or jackknifing can be performed when resampling columns of the sequence alignment.
- *Number of samples*. The number of alignments and trees to generate while resampling. A value of at least 100 is recommended.
- *Create Consensus Tree*. Choose this to create a consensus tree from the samples.
- *Sort Topologies*. Produce trees which summarise the topologies resulting from resampling. See above for more details.
- *Support threshold*. This is used to decide which monophyletic clades to include in the consensus tree, after comparing all the trees in the original set. (see Consensus Tree section above)
- *Topology Threshold*. The percentage of topologies in the original trees which must be represented by the summarizing topologies.
- *Save raw trees*. If this is turned on then all of the trees created during resampling will be save in the resulting tree document. The number of raw trees saved will therefore be equal to the number of samples.

Creating a consensus tree of existing trees

If you select a tree set document and choose "Tree" then the Consensus option will be available at the to of the tree builder options. This will create a consensus tree using the trees already in the document (no resampling will be performed) and it will be added to the tree document. A new document will not be generated.

The only option available here is the consensus support threshold.

4.6 PCR Primers (*Pro* only)

Geneious provides several operations that work with PCR Primers and DNA or hybridisation probes. PCR Primers and DNA or hybridisation probes can be designed for or tested on existing nucleotide sequences. A PCR product can be extracted from a sequence that has been

annotated with both a forward and a reverse primer. 5′ extensions consisting of restriction enzymes or arbitrary sequence may also be added to primer documents.

In addition Geneious can determine the primer characteristics for a primer sized sequence and convert it into a primer. Characteristics can also be determined for any number of primer sized selections made in the Sequence View.

To use any one of these primer operations simply select the appropriate nucleotide sequences and either select "Primers" from the Tools menu or right-click (Ctrl+click on Mac OS X) on the document(s) and select "Primers". A popup menu will appear showing the operations valid for your current selection.

4.6.1 Design Primers

The Primer Design dialog which is then displayed contains two main areas:

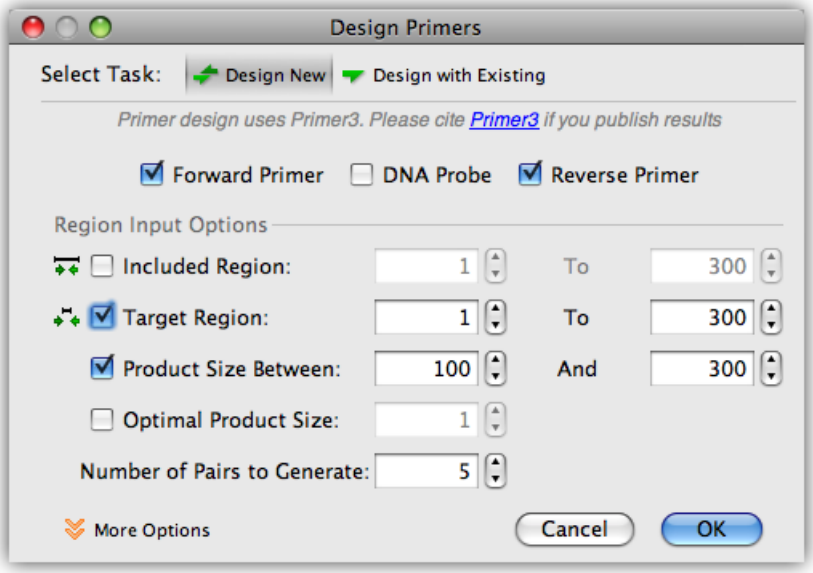


Figure 4.5: The primer design dialog

Task

Two tasks are available, "Design New" or "Design with Existing". "Design New" designs a pair of forward and reverse primers. You can specify if you wish to design with or without a matching probe. "Design with Existing" can design a partner primer to match an existing one,

for example a reverse primer for a forward or vice versa. It also allows you to design a probe to match a pair of primers.

If any documents were selected which either are primer sequences or contain primer annotations then these will be made available for selection as primers in a drop-down box. Selected sequences are treated as primer or probe sequences if they are 150bp in length or less.

Region Input Options

These options allow you to specify what part of a sequence you wish to amplify. Most options are optional and can be enabled or disabled with the associated check boxes beside them. If you have selected a region in the sequence before opening the primer dialog then this region will automatically be used for Included Region and Target Region. All of these are expressed in base pairs from the beginning of the sequence and are as follows:

- **Included Region:** Specifies the region of the sequence within which primers are allowed to fall. This must surround the target region and allows you to choose a small region on either side of the target in which primers must lie.
- **Target Region:** Specifies which region of the sequence you wish to amplify and unless the advanced options allow otherwise, the left and reverse primers must fall somewhere outside this region.
- **Product Size:** Specifies the range of sizes which the product of a primer pair can have. The product size is the distance in bp between the beginning of the left primer to the end of the reverse primer.
- **Optimal Product Size:** Specifies the preferred size of the product. Setting this will mean primer pairs that have a product size close to this will be chosen over those that do not. Warning: Setting these options can cause the primer design process to take considerably longer to complete.

The final option in this section is **Number of Pairs to Generate** which specifies how many candidate pairs of primers and DNA probes to generate and is compulsory. Setting this to 1 will give you only the primer pair which was considered best by the set parameters.

Output from Primer Design

Once the task and options have been set, click the "OK" button to design the primers. A progress bar may appear for a short time while the process completes. When complete each of the sequences will have the designed primers and probes added to them as sequence annotations. The annotations will be labelled with their rank compared to the other primers (eg. 1st,

2nd.. where 1st is the best) and what type they are (Forward primer, Reverse primer or DNA probe). Primers will be coloured green and probes red.

Detailed information such as melting point, tendency to form primer-dimers and GC content can be seen by hovering the mouse over an annotation. The information will be presented in a popup box. Alternatively, double clicking on an annotation will display its details in the annotation editing dialog.

The best way to save a primer or DNA probe for further testing or use is to select the annotation for that primer and click the "Extract" button in the sequence viewer. This will generate a separate, short sequence document which just contains the primer sequence and the annotation (so it retains all the information on the primer). In the case of the reverse primer it should be reverse complemented. When the Extract button is chosen for the reverse primer it will offer to reverse complement because the annotation runs in the reverse direction. Choose "Yes".

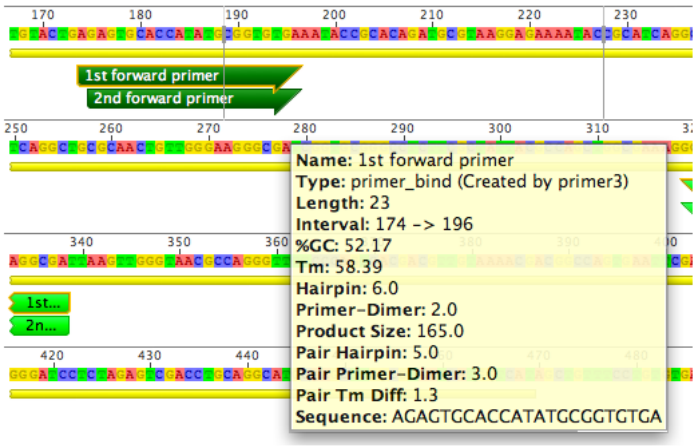


Figure 4.6: Primer design output

When no primers can be found

If no primers or DNA probes that match the specified criteria can be found in one or more of the sequences then a dialog is shown describing how many had no matches and for what reasons.

To see why no primers or DNA probes were found for particular sequences, click the "Details" button at the bottom of the dialog. The dialog will then open out to display a list of all the sequences for which no primers or DNA probes were found. For each of the sequences the following information is listed:

- Which of Forward Primer, Reverse Primer, Primer Pair and/or DNA Probe could not be found in the sequence

- For each of these, specific reasons for rejection are listed (eg. "Tm too high" or "Unacceptable product size") along with a percentage which expresses how many of the candidate primers or probes were rejected for this reason.

After examining the details you can choose take no action or continue and annotate the primer and/or DNA probes on the sequences which were successfully designed for.

4.6.2 Primer Database

The Primer Database consists of all the oligonucleotide documents that exist in your Local or Server databases. The "oligonucleotide" document type is a short nucleotide sequence representing either a primer or a probe. The text view lists the primer characteristics (Tm, GC *etc*). These properties can be shown in the document table. Tm is shown by default, but you can turn on others by right clicking on the table header.

Oligo sequences are created via one of the following methods:

- Extract a primer/probe annotation from a sequence
- Select "Sequence" → "New Sequence" from the menu and choose Primer or Probe as the type of the new sequence
- Select one or more existing primer sequences (maybe ones imported from a file) then click "Primers" → "Convert to Oligo" to transform them into oligo type sequences

If you select a target sequence and go to "Test Primers" or "Design Primers" → "Design With Existing", Geneious will find all oligo sequences in your database and offer them as options in the list of oligo sequences with no need to select them along with the target sequence before starting the operation.

The note type "Primer Info" can be used to note the fridge location *etc* of a particular primer.

4.6.3 Test Primers

Primers and probes can also be quickly tested against large numbers of sequences to see which ones the primers will bind to. By default this will only find sequences that match the primers exactly. To test primers select the target sequences you want to test for compatibility with primers and choose the same "Primers" action from the menu and go to "Test Primers" in the popup menu that appears.

Any oligo sequences that are in your local folders will be available as candidate primers but you can also select any other short sequences or sequences containing annotated primers along with the target before clicking on "Test Primers" to have these available for testing.

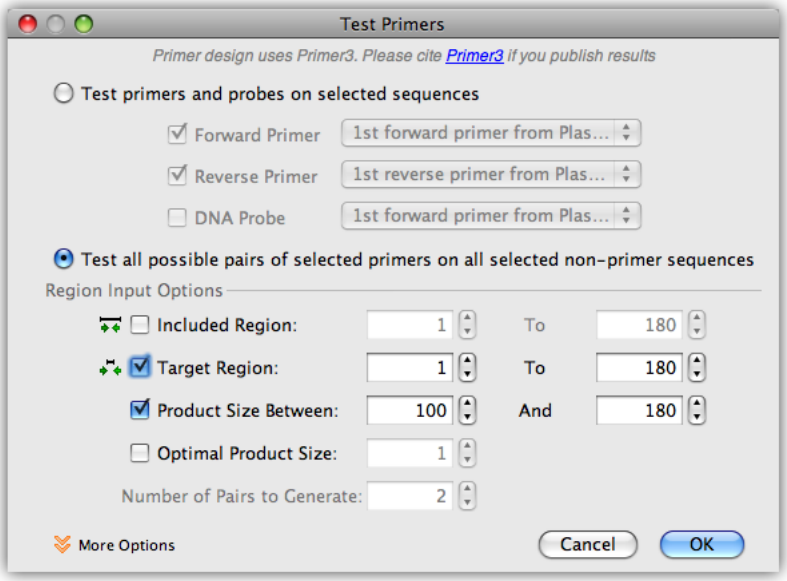


Figure 4.7: The primer test dialog

There are two ways in which Geneious can test your selection of primers and probes. The first option in the dialog tests the chosen set of primers and probes on all selected sequences. The check boxes beside each primer and probe can be used to specify if it is being tested.

The second option tests all combinations of the primers in your database along with any selected primer sequences as primer pairs on each selected target sequence. Primers that are only annotated on a larger sequence are however omitted from this process. Warning: This option has to test a large number of possible combination so can take a long time to complete.

All of the same options available for designing primers also apply to testing so if the primers are expected to bind to quite different regions of the test sequences the primer binding region may have to be extended and the target region option can be omitted.

Click the "OK" button and testing will commence. Once complete, a dialog will present the results. This dialog tells you how many of the sequences were compatible with the specified primers and probes and provides details and choices very similar to the one described in section 4.6.1. The compatible primers can be annotated onto the sequences in a similar manner to that when designing primers. Additionally if the primer sequences were not already annotated with a primer annotation they will be annotated during testing.

4.6.4 Primer Characteristics

Convert to Oligo

Geneious can convert any number of sequences that are 150 base pairs or fewer in length into primers. This operation will also determine the primer characteristics of the sequences, such as melting point. To do this, select your sequences and choose the same "Primers" action as you do with design or test, then choose "Convert to Oligo" from the popup menu that appears. If you select just two sequences you have the additional option of determining their pair characteristics. Determining the pair characteristics of two primer sequences can be used to see if two sequences can pair and how well they do so.

Characteristics for Selection

Primer Characteristics can also be determined on a selection in a larger sequence. Select a region of 150bp or less in the Sequence View and choose "Characteristics for Selection". The primer characteristics will then be added as an annotation over the exact region that was selected. This will also work on multiple selected regions in the Sequence View. Hold the Ctrl key while clicking and dragging to select multiple regions simultaneously.

4.6.5 Primer Extensions

You can add a primer extension to an existing "oligonucleotide" sequence by selecting "Primers" → "Add 5′ Extension". You can either add your own sequence, or select from available restriction sites. These extensions will not change the binding region of the primer and will be ignored when primer testing is conducted against potential target sequences.

If the primer is annotated onto a sequence following testing, the extension sequence is shown in the list of the annotation's qualifiers. If the primer or a PCR product is extracted from this annotation, the result will include the extension.

4.6.6 Advanced Options

The parameters which are used to pick primers and DNA probes are highly customisable through the advanced options section of the primers dialog. To access this, select part of a sequence for testing or designing and select "Primers" from the menu as detailed above. Now click the "More Options" button and the advanced options will appear below the standard options.

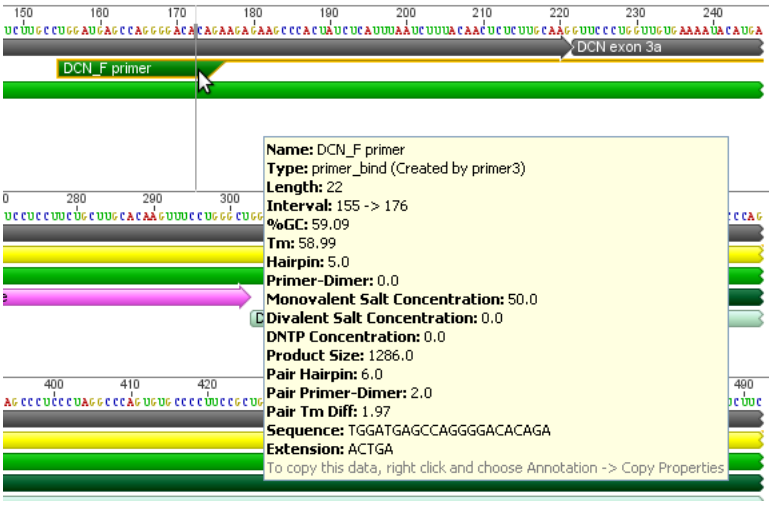


Figure 4.8: Primer annotation with extension

Additional Options

The advanced options include additional options that tell Geneious to be more lenient with how it designs and tests primers.

- **Maximum Degeneracy:** Turning this on allows Geneious to design primers which contain a certain number of ambiguities. Such a primer is called a degenerate primer. This is because the sequence actually represents more than one primer sequence. The maximum degeneracy that you specify is the maximum number of primers that any primer sequence is allowed to represent. For example, a primer which contains the nucleotide character N once (and no other ambiguities) has a degeneracy of 4 because N represents the four bases A,C,G and T. A primer that contains an N and an R has degeneracy 4 * 2 = 8 because R represents the two bases A and G.

- **Maximum Mismatches:** This is available when testing and allows you to specify a limited number of mismatches that you wish to permit between a primer and the target sequence. You can limit the position in which mismatches are allowed by clicking the "Mismatch Options" button.

- **Inverse PCR:** Enables inverse PCR which will invert the primer pair and remove the option of a target region and the ability to use a probe.

Picking Parameters

The advanced options section has two tabs which are available depending on the task you have chosen. The "Primer" section is available if one of "Forward Primer" or "Reverse Primer" is being designed or tested and "DNA Probe" is available if "DNA Probe" is being designed or tested. These two sections are quite similar; the DNA probe section has a subset of the options available in the primer section. This is because primers are usually chosen in pairs and so several options can be set for how pairs are chosen.

Most of the options are used to set absolute limits on properties of primers and probes such as melting point and GC content. Optimum values can also be specified. For details on individual options hover your mouse over them and a popup box will describe the function of the option. During testing many of the absolute limit options are disabled, however optimal values can still be set.

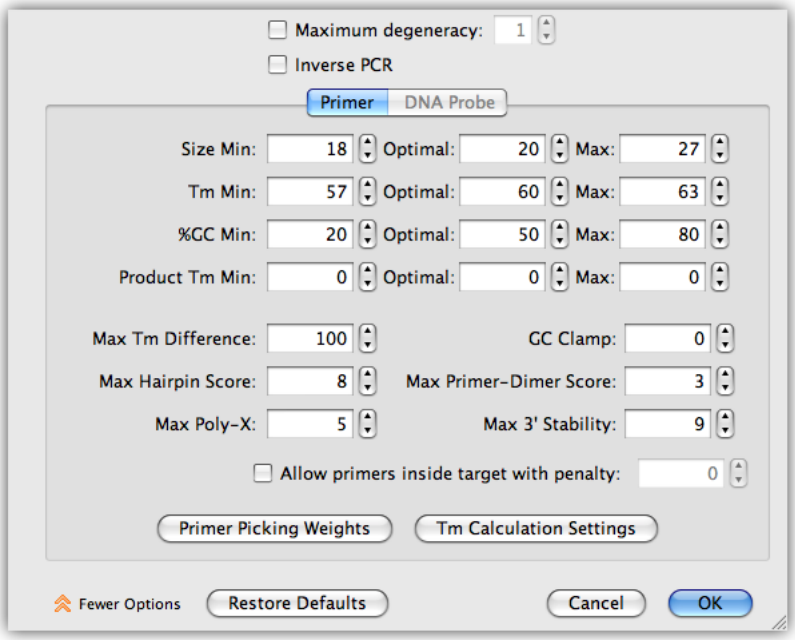


Figure 4.9: Primer design advanced options

Primer Picking Weights

At the bottom of both the advanced primer and DNA probe options there is a "Primer Picking Weights" button. Clicking this brings up a second dialog containing many more options. The purpose of all of these options is to allow you to assign penalty weights to each of the param-

eters you can set in the options. The weight specified here determines how much of a penalty primers and probes get when they do not match the optimal options. The higher the value the less likely a primer or probe will be chosen if it does not meet the optimal value.

Some of the weights allow you to specify a "Less Than" and "Greater Than". This is for options which allow you to specify an optimum score such as GC content. These weights are used when looking at primers whose value for this option falls below and above the optimum respectively. The other weights are applied no matter in which direction they vary.

For details on individual options in the Primer Picking Weights dialog, again hover your mouse over the option to see a short description.

4.6.7 More Information

The Primer feature in Geneious is based on the program Primer3 http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi.

Copyright (c) 1996,1997,1998,1999,2000,2001,2004 Whitehead Institute for Biomedical Research. All rights reserved.

If you use the primer design feature of Geneious for publication we request that you cite primer3 as:

Steve Rozen and Helen J. Skaletsky (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365-386 Source code available at http://fokker.wi.mit.edu/primer3/.

Further information on the functionality of the primer design feature can be found in the primer3 documentation available here: http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www_help.cgi. Please note that some controls have been changed, renamed or removed from Geneious, but most of the primer3 functionality is available.

4.7 Contig Assembly (*Pro* only)

Contig assembly or sequence assembly is normally used to merge overlapping fragments of a DNA sequence into a contig which can be used to determine the original sequence. The contig essentially appears as a multiple sequence alignment of the fragments. After some manual editing of the contig to resolve disagreements between fragments which result from read errors, the consensus sequence of the contig is extracted as the sequence being reconstructed.

Contig assembly is also used to align a large number of reads of the same sequence (from different individuals). This is done to find small differences between reads or SNPs (Single Nucleotide Polymorphisms). In this type of analysis the consensus sequence of the contig is

not the interesting part, the differences between fragments is. This can also be done against a known reference sequence when differences between each of the fragments and the reference are of interest.

4.7.1 Assembling a Contig

To assemble a contig firstly select all of the sequences and/or contigs you wish to assemble along with the reference sequence (if you want to use one) in the document table then click Assembly in the toolbar, in the Tools menu or in the popup menu (right-click (Ctrl+click on Mac OS) on the documents). The basic options for contig assembly will then be displayed.

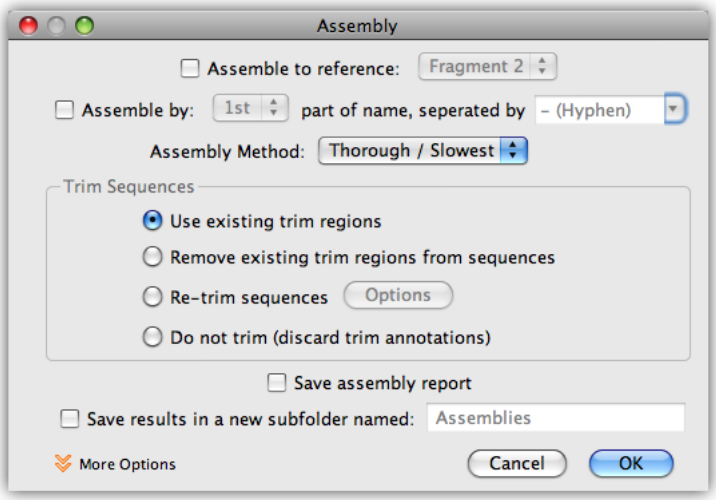


Figure 4.10: Basic assembly options

The options available here are as follows:

- **Assemble by (aka Assemble by Name):** If you have selected several groups of fragments which are to be assembled separately, you can specify a delimiter and an index at which the identifier can be found in all of the names. Sequences are grouped according to the identifier and each group is assembled separately. If a reference sequence is specified, it is used for all groups. eg. For the names A03.1.ab1, A03.2.ab1, B05.1.ab1, B05.2.ab1 etc where "A03" and "B05" are the identifiers you would choose "Assemble by 1st part of name, separated by . (full stop)"
- **Assembly method:** Specifies a trade off between the time it takes to assemble and the accuracy of the assembly. Higher sensitivity is likely to result in more reads being assembled.

- **Assemble to reference:** Select a sequence to use as the reference. See section 4.7.2.
- **Trim Sequences:** Select how to trim the ends of the sequences being assembled. See section 4.7.3.
- **Save assembly report:** Instead of displaying the results of the assembly in a dialog, the results are saved in a separate report document alongside the contig(s). This lists which fragments were successfully assembled and which contig they went in to along with a list of unassembled fragments.

Advanced Options (Click **More Options**)

- **Save results in a new subfolder named:** If selected, all results of the assembly will be saved to a new subfolder inside the one containing the fragments. This folder will always only contain the assembly results from the one most recent assembly - it creates a new folder each time it is run.
- **Alignment Options:** Penalties and scores used when aligning the fragments; these normally don't need to be changed.

Other advanced options depend on the assembly method selected. These are fully documented if you hover the mouse over them in Geneious. Thorough/slow assembly advanced options include

- **Minimum Overlap:** The minimum overlap (in nucleotides) between a sequence and any sequence in the contig required for the sequence to be included in the contig.
- **Overlap Identity:** The minimum identity (in percent) of the overlap region between a sequence and any sequence in the contig required for the sequence to be included in the contig.

Choose the options you require and click OK to begin assembling the contig. Once complete, one or more contigs may be generated. If you got more contigs than you expect to get for the selected sequences then you should try adjusting the options for assembly. It is also possible that no contigs will be generated if no two of the selected sequences meet the overlap requirements.

Note: The orientation of fragments will be determined automatically, and they will be reverse complemented where necessary.

If you already have a contig and you want to add a sequence to it or join it to another contig then just select the contig and the contig/sequence and click assembly as normal.

Click More Options in the assembly options to display the Alignment parameters. Here you can change the parameters used by Geneious when aligning fragments together. For sequences which are lower quality or contain many errors, the gap penalty should be decreased and the mismatch score should be increased.

The algorithm

Sequence assembly in Geneious uses a simple greedy algorithm which is very similar to that used in its multiple sequence alignment.

1. Determine all pairwise distances using BLAST-like search.
2. If using the thorough sensitivity option, reject pairs who's hit is not at least as long as the minimum overlap length.
3. Progressively align (using Needleman Wunsch [16] for thorough sensitivity or a quicker heuristic for less sensitive options) the highest scoring pairs (reverse complementing if necessary), appending sequences to contigs and joining contigs where necessary. Reject alignments which do not meet overlap requirements if using the thorough sensitivity setting.

4.7.2 Assembly to a reference sequence

Assembling to reference is used when you have known sequence and you wish to compare a number of reads of the same sequence with it to locate differences or SNPs. To perform assembly to a reference sequence select the sequences and the reference sequence and click Assembly. Choose the name of the sequence you wish to use as the reference in the Align to referencce option and click OK. One contig will be produced at most and this will display the reference sequence at the top of the alignment view with all other sequences below it.

See section 4.7.4 for details on identifying differences or SNPs.

When aligning to reference the sequences are not aligned to each other in any way, each of them is instead aligned to the reference sequence independently and the pairwise alignments are combined into a contig. The high, medium and low sensitivity options perform a fine tuning step after the initial assembly to make reads which overlap from the initial assembly stage align better to each other. If you just wish to use a reference sequence to help construction of the contig then you should select all sequences and the reference but choose "None" for Align to reference.

4.7.3 Trimming

Trimming low quality ends of sequences is normally performed before assembling a contig. This is because the noise introduced by low quality regions and vector contamination can produce incorrect assemblies.

The easiest way to trim sequences is at the assembly step. Select the trim options you wish to use in the Assembly options and click OK. The sequences will be trimmed and assembled in

one operation. This means you cannot view the trimming that Geneious uses before assembly is performed, but the trimmed regions will still be available and adjustable after assembly is complete.

Trimmed annotations are ignored when calculating the consensus sequence for a contig. So although the trimmed regions are visible, they do not affect the results of assembly at all.

Sequence trimming can be performed before assembly by selecting the sequences you wish to trim and selecting $Tools \rightarrow TrimEnds$. This will add "Trimmed" annotations to the sequences which are ignored in the construction of a contig. When performing "Assembly" from sequences which have been annotated in this way, select "Use Existing Trim Regions".

Trimmed annotations can also be created manually using the annotation editing in the sequence viewer. If you create annotations of type "trimmed" and save them then Geneious will treat them the same as ones generated automatically and they will be ignored during assembly. Trimmed annotations can also be modified in this way before or after assembly.

Trimming options

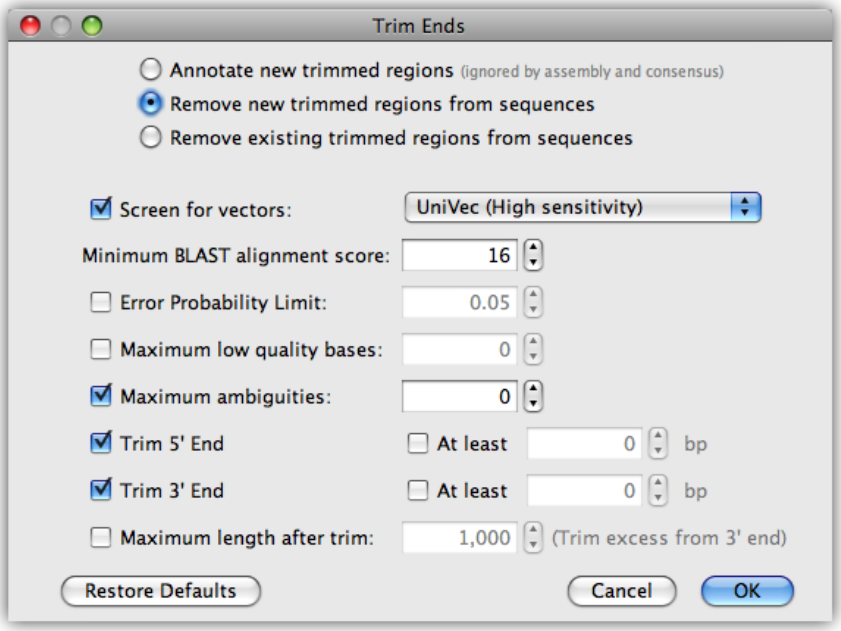


Figure 4.11: Trimming options

- **Annotate new trimmed regions:** Calculate new trimmed regions and annotate them - the

trimmed regions will be ignored when performing assembly and calculating the consensus sequence.

- **Remove new trimmed regions from sequences:** Calculate new trimmed regions and remove them from the sequence(s) completely. This can be undone in the Sequence View before the sequences are saved.
- **Remove existing trimmed regions from sequences:** This is only available when there are already trimmed regions on some of the sequences. This will remove the existing trimmed regions from the sequences permanently; no new trimmed regions are calculated.
- **Screen for vectors:** Screens the sequences against UniVec to locate any vector contamination and trim it. This uses an implementation similar to NCBI's VecScreen to detect contamination - (http://www.ncbi.nlm.nih.gov/projects/VecScreen/)
- **Error Probability Limit:** Available for chromatogram documents which have quality (confidence) values. The ends are trimmed based on these quality values using the modified-Mott algorithm (Richard Mott personal communication). This results in the longest possible un-trimmed region with an error probability no more than the limit specified.
- **Maximum low quality bases:** Specifies the maximum number of low quality bases that can be in the untrimmed region. Low Quality is normally defined as confidence of 20 or less. This can be adjusted on the Sequencing and Assembly tab of Preferences.
- **Maximum Ambiguities:** Finds the longest region in the sequence with no more N's than the maximum ambiguous bases value and trims what is not in this region. This should be used when sequences have no quality information attached.
- **Trim 5′ End and Trim 3′ End:** These can be set to specify trimming of only the 3′ or 5′ end of the sequence. A minimum amount that must be trimmed from each end can also be specified.
- **Maximum length after trim:** If the untrimmed region is longer than the specified limit then the remainder will be trimmed from the 3′ end of the sequence until it is this length.

4.7.4 Viewing Contigs

Contigs in Geneoius are viewed (and edited) in exactly the same way as alignments. There are several features in the sequence viewer which are worth taking special note of when viewing contigs:

- The consensus sequence is normally of particular interset and this is always displayed at the top of the sequence view (labeled Consensus).

- When all sequences in a contig (or alignment) have quality information attached then you can select the "Highest Quality" consensus type. This almost removes the need for manually editing the contig because this consensus chooses the base with the highest total quality at each position.
- There is a Quality color scheme which is selected by default for alignments of all chromatograms. This assigns a shade of blue to each base based on its quality. Dark blue for confidence < 20, blue for 20 - 40 and light blue for > 40. The consensus is also colored with this scheme where the confidence of a given base in the consensus is equal to the maximum confidence from the bases at that site in the alignment.
- The sequence logo graph has an option to "Weight by quality". This is very useful for identifying low quality regions and resolving conflicts.

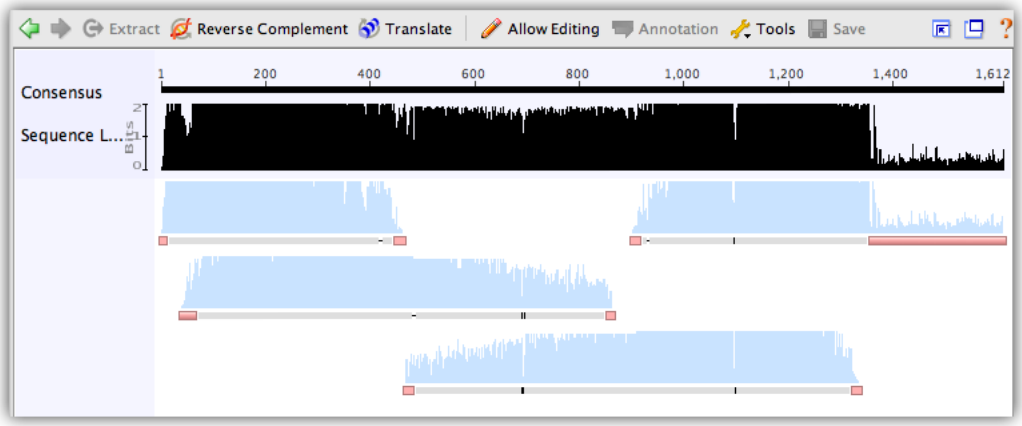


Figure 4.12: The overview of a contig

Finding disagreements or SNPs

To easily identify bases which do not match the consensus, turn on "Highlight Disagreements" in the consensus section of the sequence viewer options. When this is on, any base in the sequences which matches the consensus at that position is grayed out and bases not matching are left colored.

With this on you can quickly jump to each disagreement by pressng Ctrl+D (⌘+D on Mac OS X) or by clicking the "Next Disagreement" button in the sequence viewer option panel to the right. Each disagreement can then be examined or resolved.

You can also use this feature If you have aligned to a reference sequence and you are interested in finding differences between each sequence and the reference (or SNPs).

4.7.5 Editing Contigs

Editing a contig is exactly the same as editing an alignment in Geneious. After selecting the contig, click the Edit button in the sequence viewer and you can modify, insert and delete characters like in a standard text editor.

Editing of contigs is done to resolve conflicts between fragments before saving the final consensus. The normal procedure for this is to look through the disagreements in the contig (as described above) and change bases which you believe are bad calls to be the base which you believe is the correct call. This is often decided by looking at the quality for each of the bases and choosing the higher quality one. Geneious can do this automatically for you if you use the "Highest Quality" consensus.

Bases in the consensus sequence can also be edited which will update every sequence at the corresponding position to match what is set in the consensus.

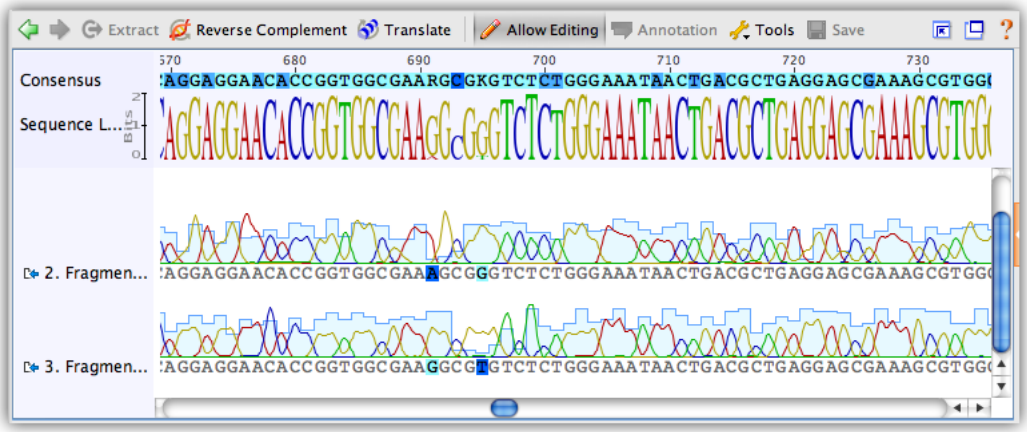


Figure 4.13: Highlight disagreements and edit to resolve them

4.7.6 Saving the Consensus

Once you are satisfied with a contig you can save the consensus as a new sequence by clicking on the name of the consensus sequence in your contig and clicking the Extract button.

4.8 Results of analysis

All analysis results are deposited in the currently selected folder. If no local folder is selected then you will be prompted for a local folder. This applies to sequence alignments, phylo-

genetic trees, sequence translations, reverse complements and extraction of sequences. Once generated, analysis results can be dragged to another location if desired.

Chapter 5

Custom BLAST (*Pro* only)

Custom BLAST allows you to create your own custom database from either FASTA files or sequences in your local folders, and BLAST against it.

5.1 Setting Up

The Custom BLAST plugin requires access to NCBI BLAST binary files.

5.1.1 Setting up the Custom BLAST files yourself

If you want, you can download or otherwise acquire the NCBI BLAST binary files outside of Geneious. You can download them from here:

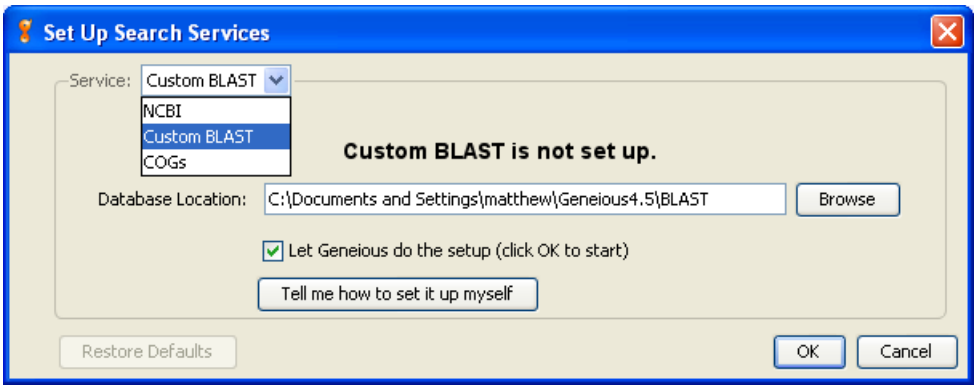
`ftp://ftp.ncbi.nih.gov/blast/executables/LATEST`

Choose the appropriate file for your operating system, download and extract it. You will need to let Geneious know where to look for the files once you have done this. To do this, go to $Tools \rightarrow Add/Remove\ Databases \rightarrow Set\ Up\ Search\ Services$ and select Custom BLAST from the Service drop-down box. Enter your data location or click "Browse" to browse to the location of the files.

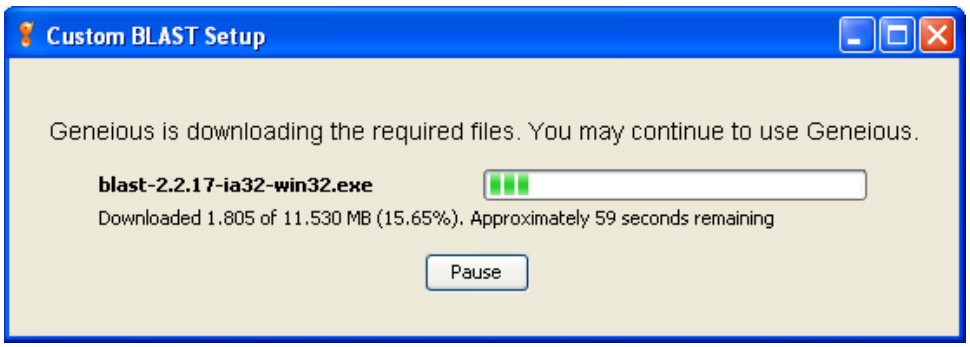
5.1.2 Setting up the Custom BLAST files through Geneious

Geneious provides a download manager to help you download and extract the Custom BLAST files. To use it, go to Tools $Tools \rightarrow Add/Remove\ Databases \rightarrow Set\ Up\ Search\ Services$ and select Custom BLAST from the Service drop-down box. Make sure "Let Geneious do the

setup" is checked. Then click "OK". After a few seconds the compressed file containing all the files needed to run Custom BLAST will start downloading. You can click "Pause" to pause the download. You can add and search Custom BLAST databases as soon as it has finished downloading and extracting. If you shut down Geneious with the file partially downloaded, you will need to start downloading it again from the beginning.



(a) Setup Options


Custom BLAST Setup
Geneious is downloading the required files. You may continue to use Geneious.
blast-2.2.17-ia32-win32.exe
Downloaded 1.805 of 11.530 MB (15.65%). Approximately 59 seconds remaining
Pause

(b) Downloading

Figure 5.1: Setting Up Custom BLAST

5.1.3 Adding Databases

Now that you have set up the executables, it is time to add databases to your BLAST.

Adding FASTA databases

To create a database from the sequences in a FASTA file, go to *Tools* → *Add/Remove Databases* → *Add Sequence Database* and select Custom BLAST from the Service drop-down box. Choose to "Create from file on disk" and then click "Browse" to navigate to the FASTA file that contains

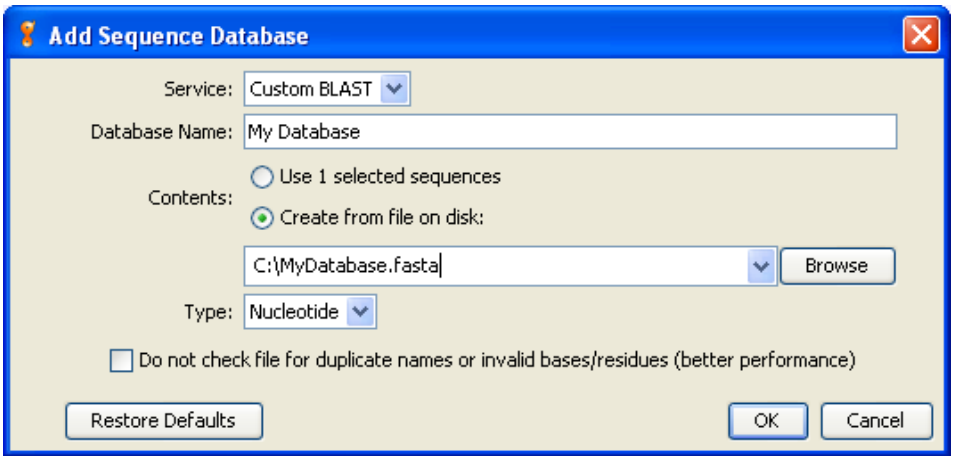


Figure 5.2: Adding a FASTA database

the sequences you want to BLAST. Enter a name for the database and click OK. There are two requirements for a FASTA file to be suitable for creating a database from:

- The FASTA file must contain only the same types of sequence (i.e. Nucleotide or Amino Acid)
- The sequences in the FASTA file must all have unique names

If the file meets these requirements it will be added as a database, otherwise you will be informed of the problem.

Creating a database from local documents

To create a BLAST database from sequences in your local documents folders, first select the documents that you want. Then go to $Tools \rightarrow Add/Remove\ Databases \rightarrow Add\ Sequence\ Database$ and select Custom BLAST from the Service drop-down box. Enter a name for the database,and click OK.

5.1.4 Using Custom BLAST

Once you have added one or more databases, they will appear under Custom BLAST in the Sequence Search database drop down. These can be used in exactly the same way as the NCBI BLAST ones.

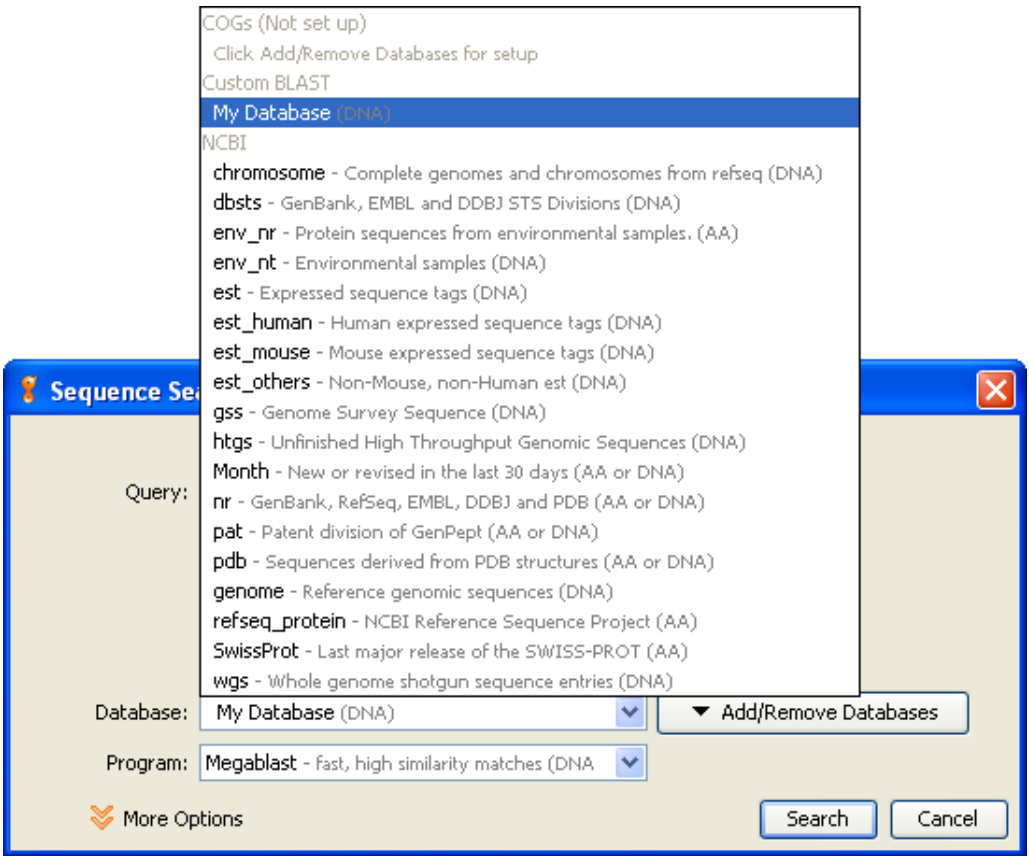

COGs (Not set up)
Click Add/Remove Databases for setup
Custom BLAST
My Database (DNA)
NCBI
chromosome - Complete genomes and chromosomes from refseq (DNA)
dbsts - GenBank, EMBL and DDBJ STS Divisions (DNA)
env_nr - Protein sequences from environmental samples. (AA)
env_nt - Environmental samples (DNA)
est - Expressed sequence tags (DNA)
est_human - Human expressed sequence tags (DNA)
est_mouse - Mouse expressed sequence tags (DNA)
est_others - Non-Mouse, non-Human est (DNA)
gss - Genome Survey Sequence (DNA)
htgs - Unfinished High Throughput Genomic Sequences (DNA)
Month - New or revised in the last 30 days (AA or DNA)
nr - GenBank, RefSeq, EMBL, DDBJ and PDB (AA or DNA)
pat - Patent division of GenPept (AA or DNA)
pdb - Sequences derived from PDB structures (AA or DNA)
genome - Reference genomic sequences (DNA)
refseq_protein - NCBI Reference Sequence Project (AA)
SwissProt - Last major release of the SWISS-PROT (AA)
wgs - Whole genome shotgun sequence entries (DNA)
Sequence Se
Query:
Database:
My Database (DNA)
Add/Remove Databases
Program:
Megablast - fast, high similarity matches (DNA
More Options
Search
Cancel

Figure 5.3: Searching a Custom BLAST database

Chapter 6

COGs BLAST(*Pro* only)

COGs BLAST allows you to BLAST against the COGs database (http://www.ncbi.nlm.nih.gov/COG/). Geneious will BLAST your sequence against the COGs database, identify which COG the sequence is most likely to reside in, and give you information about the COG.

6.1 Setting Up

To set up the COGs database, you first need to set up Custom BLAST on your computer (see the section on Custom BLAST). Once you have set up Custom BLAST, you need to set up the COGs database files.

6.1.1 Downloading the COGs BLAST files yourself

If you want, you can download or otherwise acquire the COGs BLAST database files outside of Geneious. You can download them from here:

(ftp://ftp.ncbi.nih.gov/pub/COG/COG/).

The files you need are:

- myva
- myva=gb
- whog

Save these files to a local folder. Now go to *Tools* → *Add/Remove Databases* → *Add Sequence Database* and select Custom BLAST using the Service drop-down box. Choose to "Create from file on

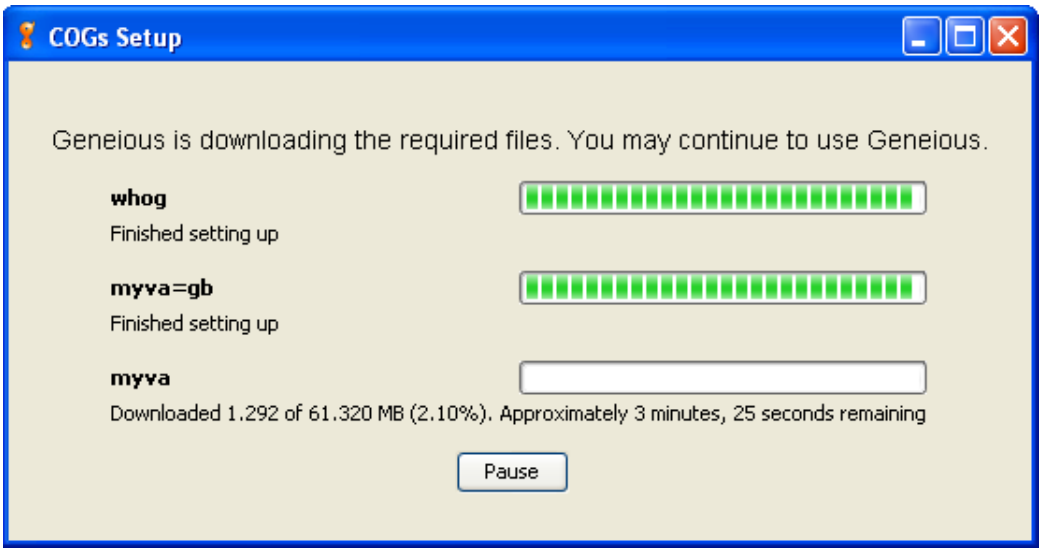

COGs Setup
Geneious is downloading the required files. You may continue to use Geneious.
whog
Finished setting up
myva=gb
Finished setting up
myva
Downloaded 1.292 of 61.320 MB (2.10%). Approximately 3 minutes, 25 seconds remaining
Pause

Figure 6.1: The Cogs BLAST Download Manager

disk" and click "Browse". Navigate to the file myva and click OK (make sure that the protein database option is checked). Now copy the other two files that you downloaded into the data folder inside your Custom BLAST folder.

6.1.2 Downloading the COGs BLAST databases through Geneious

Geneious provides a download manager to help you download and set up the COGs BLAST database. To use it, go to Tools $Tools \rightarrow Add/Remove\ Databases \rightarrow Set\ Up\ Search\ Services$ and select COGS BLAST from the Service drop-down box. Make sure "Let Geneious do the setup" is checked. Then click "OK". After a few seconds the compressed file containing all the files needed to run COGS BLAST will start downloading

You can click "Pause" to pause the download. Once all the files have finished downloading aned setting up, you will need to close the dialog. If you shut down Geneious with a file partially downloaded, you will need to start downloading it again from the beginning. Files completely downloaded will not need to be downloaded again.

6.2 BLASTing COGs

Select any sequence in the document table, right click it, and select "Sequence Search". Select the COGS database from the database drop-down box and Geneious will give you several options for your blast (see Figure 6.2). Number of hits to fetch allows you to fetch results for

the best *n* hits for your sequence. You can choose to download COGs sequence from NCBI (with full annotations) or to load them without annotations from the COGs database file. Finally you have the option of retrieving the sequences for your hits, the entire COG for each hit, or to just display information about the hits. Once you have made your choices, click OK. If you have selected a Nucleotide sequence, Geneious will give you options to translate it at this point.

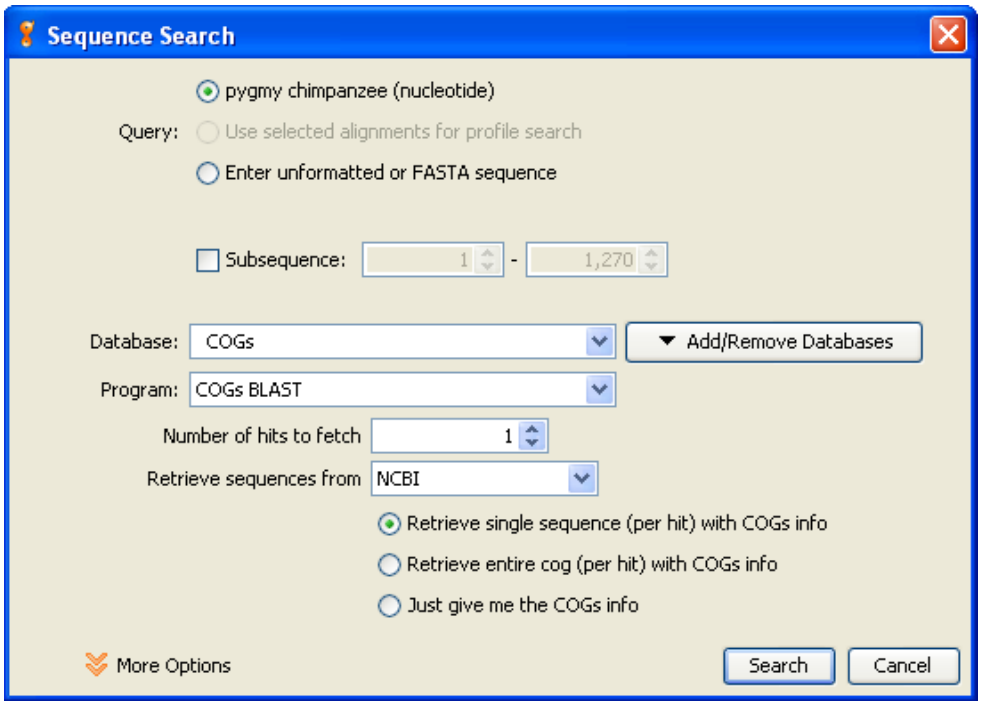


Figure 6.2: Configuring a COGs BLAST

Chapter 7

Pfam (*Pro* only)

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. The data for Pfam is taken from sequences in UniProt. Pfam can be found online at the following locations:

- Sanger Institute (UK)
- Washington D.C. (USA)
- Karolinska Institutet (Sweden)
- Institut National de la Recherche Agronomique (France)

7.1 Setting up the Pfam databases

At the time of release of Geneious 3.5, there was no public online interface to the Pfam database, (although there is one in the works at the Sanger Institute). For this reason, if you want to search the Pfam databases, you will need to download them first. As of Pfam 22 (July 2007) the subset of the Pfam databases used by Geneious totalled about 4GB in size, so it is recommended you download them somewhere with a fast connection.

You can use Geneious to search five of the Pfam databases:

1. **Pfam-A.seed** (29 MB) contains records on the manually curated domains in Pfam-A and the seed alignment (alignment of a representative subset of all occurrences of this domain in UniProt sequences) for each domain
2. **Pfam-A.full** (392 MB) contains records for the manually curated domains in Pfam-A and the full alignment (alignment of all occurrences of this domain in UniProt sequences) for each domain

3. **Pfam-B** (59 MB) contains records for the automatically generated domains in Pfam-B taken from PRODOM
4. **Pfam-C** (69 KB) contains records for Pfam clans (families of similar domains)
5. **swisspfam** (132 MB) contains data on the domain architecture of UniProt sequences.

7.1.1 Downloading the Pfam databases yourself

If you want, you can download or otherwise acquire the Pfam databases outside of Geneious. You will need to let Geneious know where to look for the files once you have done this. To do this, select the Pfam service. Click the "Change Database Location" and browse to the location of the databases.

7.1.2 Downloading the Pfam databases through Geneious

Geneious provides a download manager to help you download the Pfam files. To use it, select the Pfam Service. Click the "Let Geneious do it" button. Then click the "Start" button. After a few seconds the first database will start downloading. You can click "Pause" to pause the download. You can search a database as soon as it has finished downloading and its contents have been verified. If you shut down Geneious with a file partially downloaded, you will need to start downloading it again from the beginning.

The Pfam databases total around 4 GB in size, most of which comes from Pfam-A.full. If your internet connection is slow or you have a low data cap you may want to download the databases elsewhere, and then transfer them to your computer. You may also consider downloading all databases except Pfam-A.full.

7.2 Pfam Document Types

There are three special document types used for Pfam data:

Pfam sequence documents are based on UniProt sequences. They contain all the information from the UniProt sequence, plus information on the Pfam domains in the sequence. You can view the domains as annotations in the sequence view, or on their own from the domain view.

Domain documents contain information about Pfam A full, Pfam A seed and Pfam B domains. This includes general information about the domain, references (visible in the reference view) and the alignment for the domain.

Clan documents contain information about a clan, including general information, refer-

ences (visible in the reference view) and a list of the domains which are members of this clan.

7.3 Pfam Operations

There are a number of special operations available to Pfam documents and UniProt sequences. To take advantage of these operations, you will need to have the Pfam databases set up.

The following Pfam operations are available:

- **Create Pfam Sequence** creates a Pfam sequence document from a UniProt sequence. You can view the domain information in a Pfam sequence document using the Domain Viewer. This operation can take a long time.
- With **Find Similar Sequences** you can search and create documents for sequences in UniProt which match the domain architecture of your Pfam sequence document, ie they have the same domains in the same places. This operation can take a long time.
- **Get Domains in Sequence** creates a domain document for every domain in a Pfam sequence document.
- If your domain document is a member of a Pfam clan, you can use **Get Clan** to get a document representing that clan.
- **Get Domains in Clan** will do the opposite, ie get documents representing each domain in a clan.
- If your domain document contains the seed alignment for the domain, you can use **Get Full Alignment** to get a domain document with the full alignment.
- Conversely, you can use **Get seed alignment** to get a domain document with the seed alignment only from a domain document with the full alignment.
- **Get Full Sequences** will return the full UniProt sequence documents from which the sequences in the alignment in a domain were extracted.
- **Get Full Sequence** will return the full UniProt sequence document from which a sequence taken from an alignment in a domain was extracted.

Chapter 8

Smart Folders (*Pro* only)

Smart folders are a new feature of Geneious that allow you to separate relevant data from extraneous search results retrieved by an agent.

Smart Folders are created from within the "Create Agent" dialog. To open the Create Agent dialog, choose the "Agents" button from the toolbar, and then select "Create" from the agents dialog. Choose a folder for the agent, or create a new one, and make sure that the "Make destination folder a smart folder" checkbox is checked.

When a folder is turned into a smart folder, it is given a subfolder called "reject". At first, all the documents delivered by the agent will be put in this folder. Drag the documents that you want to keep into the main folder, and future documents delivered by the agent will be compared to the accepted and the reject documents, and stored in one or other of the two folders appropriately. Make sure that you leave documents in the reject folder, as smart folders need negative examples to build an accurate comparison model. Note that unread documents in the main folder will not be compared, while all documents in the reject folder will be.

Chapter 9

Geneious Education (*Pro* only)

This feature allows a teacher to create interactive tutorials and exercises for their students. A tutorial consists of a number of HTML pages and Geneious documents. The student edits the pages and documents to answer the tutorial questions, and then exports the tutorial to submit for marking.

9.1 Creating a tutorial

The backbone of Geneious Tutorials are the HTML documents. Simply create your documents, and place them together in a folder. If you make a page called "index.html", it will be treated as the main page. Geneious will follow all hyperlinks between the pages, and external hyperlinks (beginning with http://) will be opened in the user's browser. If you want to include figures and diagrams in the pages, just put the image files in the folder and reference them with `<img>` tags like a normal HTML document *(supported image formats are GIF, JPG, and PNG)*.

If you want to include Geneious documents in your tutorial, simply place them in the folder as above and they will automatically be imported into Geneious with the tutorial. If you want to link to them from the tutorial pages, create a hyperlink pointing to the file in the HTML document. For example, to create a link to the file sequence.fasta in your tutorial folder, use the HTML `<a href="sequence.fasta">click here</a>`. To open more than one document from a link, separate the filenames with the pipe *(|)* character, for example `<a href="sequence.fasta|sequence2.fasta">click here</a>`. Note that geneious files must contain only one document to be imported automatically with the tutorial.

You can add a short one-line summary by writing your summary in a file called *"summary.txt"* (case sensitive) and putting it in the tutorial folder. Make sure that the entire summary is on the first line of the file, as all other lines will be ignored.

Once you have all your files together, put the contents of the folder in a zip file with the exten-

sion *.tutorial.zip*. Be careful not to put subfolders in your zip file, as these are not supported.

9.2 Answering a tutorial

Import the tutorial document into Geneious (use "File" → "Import" → "From file"). The tutorial document and any associated geneious documents will be imported into the currently selected folder. The tutorial itself will be displayed in the help pane on the right hand side of the Geneious window. If you accidentaly close the help pane, you can display it by choosing Help from the Help menu.

If the tutorial requires you to enter answers, click the edit button at the top of the tutorial window and type your answer in to the space provided. Click the save button when you are done.

If the tutorial has a link to a Geneious document, when you click the link the document will be opened in the document viewer. Any changes you make to this document will be preserved when you export the tutorial.

When you have finished the tutorial, export it by selecting the tutorial document and choosing "File" → "Export" →"Selected Docuemnts" from the main menu. Make sure that "Geneious Tutorial File" is selected as the filetype, and then give it a name and click Export.

Chapter 10

Collaboration (*Pro* only)

Collaboration allows *Geneious pro* users to share the products of their research and work with each other. Based on an open Internet protocol called *XMPP* or *Jabber*, it allows you to maintain a list of contacts, so that you see who is online when you sign on yourself. You can then share documents with your online contacts, and browse and work with their documents in return. The list of contacts is stored on the server, so you can easily access an account including its contacts both at work and on your private computer.

Collaboration can work with any existing Jabber service, such as Google Talk, but we recommend using the Geneious default, `talk.geneious.com`.

You can even access several Jabber accounts at the same time, which is particularly convenient if you wish to set up and run your own Jabber server (section 10.5.3).

This chapter shows you how to:

- Create a new collaboration account
- Search for, and add contacts to your account
- Share local folders with your contacts
- Search your contacts as you would an online database
- Set up and run your own Jabber server

10.1 Managing Your Accounts

When you start Geneious you will see the empty Collaboration service in the Services Panel and the Collaboration menu at the top. You can open the Add New Account dialog by either

right-clicking (Ctrl+click on Mac OS X) on Collaboration in the Services Panel and clicking, 'Add New Account' in the popup menu, or by selecting the same option from menu at the top.

10.1.1 Add New Account

In this dialog you are given the options of creating a new account on the server or entering the details for an existing account (e.g. if you want to access an account from an additional computer). If you choose to create a new account Geneious will attempt to automatically register your account on the server at the end of this process.

Figure 10.1: Add New Account dialog box

Choose a username and password now. Enter your password twice for a new account.

You can also optionally add an email address. Biomatters will need this if you require support regarding e.g. reset of password or deletion of accounts.

More Options You can change some of the defaults for new and exiting accounts:

- *Account Name* is the name displayed in the Services Panel for this account. It defaults to your username if nothing is entered
- *Server* is the server your account connects to (default: `talk.geneious.com`).
- *Jabber Service Name* is required by some other Jabber service providers, such as Google Talk. Don't enter anything here unless you know what you are doing.
- *Port Number* for Jabber servers running on a non-standard port (default: 5222).

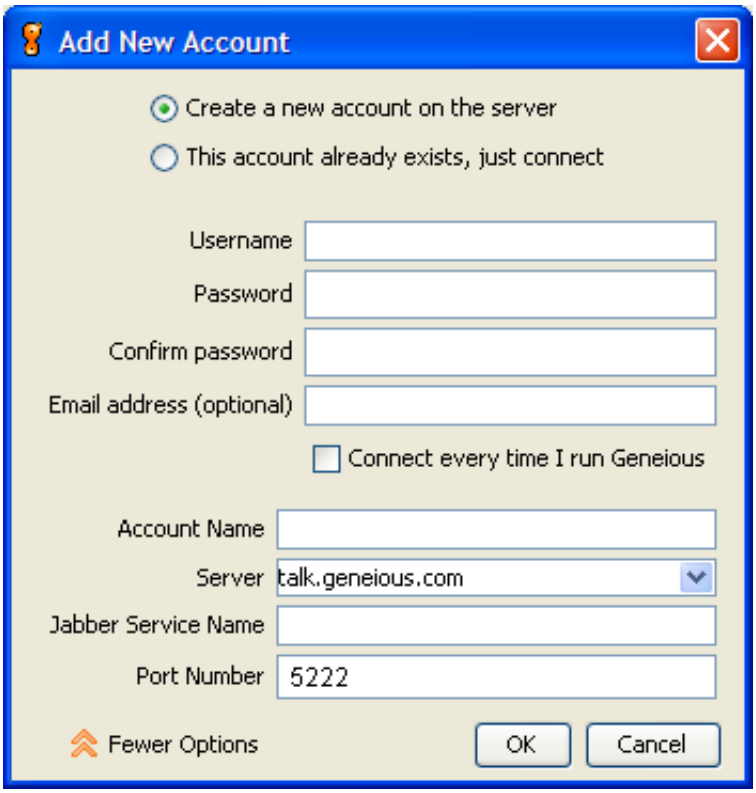


Figure 10.2: Add New Account dialog box with More Options

10.1.2 Edit Account Details

This option (from the Collaboration menu, or your account's context menu) allows you to change the configuration you made when creating the account. If you change your password, Geneious will attempt to change it on the server the next time you connect. For this purpose, Geneious internally remembers your previous password as well, so that it can still connect if you have entered your new password while disconnected.

10.1.3 Connect/Disconnect

As all other collaboration-related commands, options for connecting to or disconnecting from your account are available both in the Collaboration menu and your account's context meu (right-click, or on Ctrl+click on Mac OS X, on your account).

10.1.4 Delete Account

This option deletes your account configuration from Geneious. Currently, there is no option for deleting an account on the server.

10.2 Managing Your Contacts

Once you have an account and are connected you can start adding contacts. You will not be able to add contacts while an account is disconnected. Also, you will not be able to see a contact's online status until that contact has approved your request to do so.

10.2.1 Add Contact

Select your account in the Services Panel and choose Add Contact from the menu at the top or right-click (Ctrl+click on Mac OS X) on your account in the Services Panel and choose the same option.

You will see a simple dialog with one field, Jabber ID. A Jabber ID looks like an email address and has a similar function: It uniquely identifies some other Geneious users account. You can enter a contact's Jabber ID directly into this field if you know it. To see your own Jabber ID hover your mouse over your account in the Services Panel and it will appear in a tool-tip.

If the server supports it, you should also see a "Search For Contact" link. Click this to go to the next dialog.

Figure 10.3: Add Contact dialog box

Here you will see a box for a search string, and some checkboxes indicating what you are searching on. Enter all or part of the name or email of the contact you want and click the Search button. If any rows are returned in the results table you will be able to select one or entries and add them as contacts.

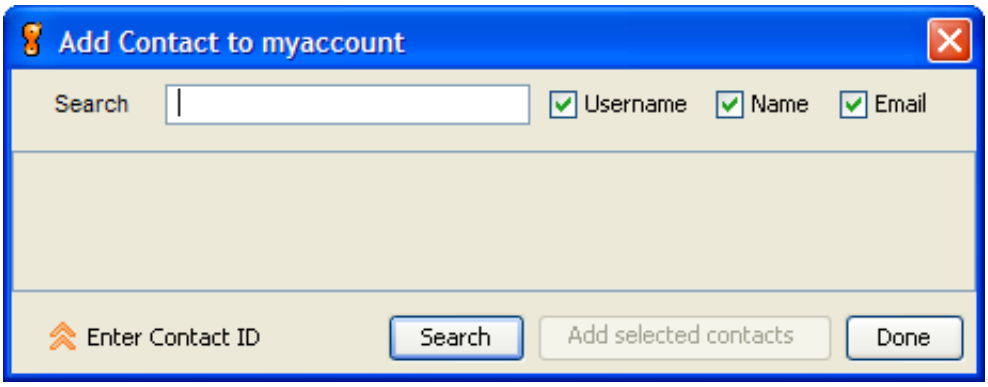


Figure 10.4: Add New Contact dialog box in searching mode

Your new contact will appear immediately in your contact list, however you will not be able to tell whether your new contact is online until they accept you as a contact. Similarly you will occasionally see a dialog box pop up asking you, "Allow user.name@talk.geneious.com as contact?" This is another Geneious user attempting to add you as a contact in this manner.

Your contact will appear grey in your contact list when they are offline. If your contact is online, they will appear blue. A contact online in Geneious will have the orange Geneious 'G' behind them. A contact online in some other program, like a chat client, whill have a speech bubble behind them.

10.2.2 Rename Contact

This option allows you to change the name that you know another contact by. This is the name the contact will appear under in the contact list and in chats; it is only visible to you.

Figure 10.5: Rename Contact dialog box

10.2.3 Remove Contact

If you no longer wish to share documents with a contact, you can remove that contact by right-clicking (Ctrl+click on Mac OS X) the contact in the Services panel and selecting "Remove Contact...". This deletes you from their contact list as well. If you find that a contact has disappeared from your list, this may be the reason.

10.3 Sharing Documents

Select one of your local folders. Select Share Folder from the File menu. Alternatively right-click (Ctrl+click on Mac OS X) on a local folder and select the same option.

- If you share a folder all documents in that folder are shared.
- If you share a folder all sub-folders of that folder are shared.
- If you share a folder it is available to all your contacts. In the future, Geneious may support per-account options for sharing your documents, or even organize contacts into groups so that you can share your documents with specific groups only.

10.4 Browsing, Searching and Viewing Shared Documents

Folders that your contacts have shared will appear beneath that contact just as they do in your contact's own Services panel. You can browse these folders as you do your local folders. You

can also search a shared folder just as you can a local one.

Additionally, you can search all of a contact's shared documents by clicking on the contact itself and then conducting the search. You can also search all the shared documents of all of an account's contacts by clicking on the account and conducting the search. Agents can be set up on shared folders, contacts and accounts.

You cannot search, browse or run or set up agents on a contact that is currently offline.

When you first view your contact's documents in the Document Table, the documents you see are only summaries. To view the whole document, select the summary(s) of the documents(s) you would like to view and the click the "Download" button inside the document view or just above it. There are also "Download" items in the File menu and in the popup menu when document summary is right-clicked (Ctrl+click on Mac OS X). The size of these files is not displayed in the Documents Table. You can cancel the download of document summaries by selecting "Cancel Downloads" from any of the locations mentioned above.

10.5 Chat

You can either chat with a single contact, or invite several contacts to join you in a new chat.

10.5.1 Chatting with One Contact

To start chatting with a particular contact (who may be online using Geneious or another chat client which uses the Jabber protocol), click on that contact and select "New Chat Session..." either from the Collaboration menu or from the popup menu (right-click on the contact, or Ctrl+click on Mac OS X). Type your messages into the text field at the bottom of the window that pops up, and click *Send* or press the Enter key to send.

10.5.2 Chatting with Multiple Contacts

Starting a Chat Session with Multiple Contacts

To invite several contacts to join you in a new chat session, click on your account (not the contacts) and then select "New Chat Session..." from either the Collaboration menu or the context menu (right-click on the account, or Ctrl+click on Mac OS X). Select the online contacts which you want to invite (you can select a range by Shift+clicking, or add contacts to the selection by Ctrl+clicking). Click *Invite* to create this new chat session.

Accepting or Declining an Invitation to Chat

When one of your contacts invites you to chat, a dialog will appear, asking you to accept or decline the chat invitation. Clicking *Accept* will open a chat window that will allow you to chat with the contact who invited you, and with all other contacts that were invited. If you decline that invitation and enter a reason (optional), this reason will be displayed to everyone in the chat.

Sending and Viewing Messages in the Chat

The chat window displays your own and your contacts' previous messages. You can enter new messages in the field at the bottom. These messages will only be sent and become visible to your contacts once you click *Send* or press the Enter key.

To leave the chat, simply close the Chat Window.

10.5.3 Setting up and running your own Jabber server

Setting up your own Jabber server is simple and means that your documents will never leave your local network. This means that you will not have any problems with firewalls, achieve much greater download speeds, and it provides an extra security layer for the confidentiality of your documents, in case it is not sufficient for you that the communication with our Jabber server is encrypted, and that we do not log or share your data.

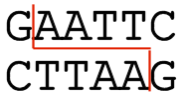
If you wish to set up and run your own Jabber server, we recommend using Wildfire from Ignite Realtime [http://www.igniterealtime.org/projects/wildfire/index.jsp/] which is available for free under the GNU General Public License. [http://www.gnu.org/copyleft/gpl.html] Install and start the server on one computer, and then enter that computer's name or address in the *Server* field under *More Options*, when creating a new account.

Please note that Biomatters cannot provide any further support for setting up and managing your Jabber server, except possibly under a contracting agreement.

Chapter 11

Cloning (*Pro* only)

Restriction Enzymes cut a nucleotide sequence at specific positions relative to the occurrences of the enzyme's *recognition sequence* in the sequence. For example, the enzyme EcoRI has the recognition sequence GAATTC and cuts both the strand and the antistrand sequence after the G inside the recognition sequence[1], leaving a single-stranded overhang (*sticky end (overhang)*):



The cloning features in Geneious allow you to identify candidate Restriction Enzymes[2] for your experiments and to determine *in silico* where they would cut your nucleotide sequences and which fragments they would produce. It also lets you ligate fragments and insert a fragment into a vector. If you select a nucleotide sequence, restriction analysis is available under the menu item Tools / Restriction Analysis, and in the context menu (right-click on the sequence, or Ctrl+click on Mac OS X):

- *Find Restriction Sites...* allows you to specify an arbitrary candidate set of restriction enzymes and the desired number of matches (so that you can e.g. identify enzymes that cut only once or twice), as well as a region enzymes may not cut within. After running the analysis, the position of the matching enzymes' recognition sequence and the sites where they cut will be visible on the sequence as annotations, and you will be able to see a table of all fragment start and end positions and their lengths, and of all restriction enzymes involved. These tables can be exported as .csv files for subsequent processing with other software such as e.g. Microsoft Excel.

[1] Like many restriction enzymes EcoRI is methylation dependent and cuts only if the second A in the recognition sequence is not methylated to N6-methyladenosine.

[2] The restriction enzyme information included in Geneious was obtained from **Rebase** [18], available for free at http://rebase.neb.com.

- *Digest into fragments...* allows you to generate the actual fragments that would be created in a digestion experiment using restriction enzymes.. When running a digestion experiment, you can choose to either use the restriction sites already annotated to the sequences (or a subset that corresponds to only some specific enzymes), or you can let Geneious determine the cut sites for any candidate enzymes. The latter option finds the cut sites for the candidate enzymes and generates the fragments in a single step.

- *Ligate Sequences...* lets you ligate two or more fragments, with or without overhangs

- *Insert into Vector...* allows you to choose a digested fragment or a sequence with two restriction site annotations to use as an insert, and insert them into a vector (circular sequence). Geneious can do the work of working out what cut sites on the vector are compatible with the overhangs on the insert, with some additional information from you.

The following sections explain the more complicated operations in a little more detail.

11.1 Find Restriction Sites

The option *Find Restriction Sites...* from the Tools / Cloning menu or the context menu allows you to find and annotate restriction sites on a nucleotide sequence. You can configure the following options (Figure 11.1):

- *Candidate Enzymes* lets you select a set of restriction enzymes from which you want to draw the ones to use in the analysis. This will always include the option to use all known commercially available restriction enzymes, but if your search index is intact then all restriction enzyme set documents from your local database will also be listed (see below for how to create such a document).

- *Minimum effective recognition sequence length* lets you filter the candidate enzymes to include only ones whose recognition sequence has a given minimum effective length. For example, EcoRI's recognition sequence is 6 nucleotides long (GAATTC). The *effective* length takes ambiguities into account, so that e.g. the sequence YS only has an effective length of 1; it is a better measure for the expected number of hits in a random sequence of fixed length, because YS matches CC, CG, TC and TG: On a random sequence with uniform nucleotide distribution it would match approximately once every nucleotide, as would a recognition sequence of length 1; hence, the *effective* length of YS is 1.

- *Only include enzymes that match X to Y times* lets you filter the results once the restriction sites have been identified. If checked, this option will discard all restriction sites for enzymes whose recognition sequence matches less than X or more than Y times. If you set X to be 0, when this operation is complete, it will report which candidate enzymes matched 0 times.

- *Exclude enzymes cutting between residues* lets you annotate only enzymes which do not cut within a certain range.

- If you select to show *More Options*, a table of all enzymes in your candidate set (filtered by the effective recognition sequence length constrained, when active) will be displayed. Only the enzymes selected in this table will be considered in the analysis; initially, all rows are selected. You can click on the column headers to sort the table ascending or descending by that column, and you can Shift+click and Ctrl+click to select a range of rows and to toggle the selection of a row, respectively.

- If not all candidate enzymes are currently selected (because of a recognition sequence length constraint, or because you have selected a subset of the table rows yourself), you can save the currently selected enzymes into a separate document by clicking *Save Selected Enzymes*. The document will be created in the current folder in your local database, and this set will then beavailable in the *Candidate Enzymes* option in this and all future analyses until the document is deleted. You can choose a custom name for the document, such as *Lab Fridge* or *Enzymes in pBlueScript II SK(+) multiple cloning site*.

After configuring your options, click *OK* to start the analysis and annotate the restriction sites on the sequence, or *Cancel* to abort.

11.2 Digest into fragments

The option *Digest into fragments...* from the Tools / Cloning menu or the context menu allows you to generate the nucleotide sequences that would result from a digestion experiment. You can digest multiple nucleotide sequences at a time. If the digestion results in overhangs, these will be recorded as annotations on the fragments.

- If you have selected only one nucleotide sequence document and it has annotated restriction sites, you can select *Digest using Annotated cut positions* to cut the document on these sites. When this option is selected, the options to filter the enzymes by their effective recognition sequence length or number of hits are disabled. However, if you select a subset of the enzymes under *More Options*, only the cut sites from these enzymes will be considered; this can easily be used for the same effect by sorting by columns and then selecting a range of rows, in the rare cases when it is needed.

- Otherwise, if you select *Digest using Enzyme Set*, the digestion operation includes finding the restriction sites first (but without generating the annotations). Therefore, the options are the same as for *Find Restriction Sites...*, which is discussed in section 11.1.

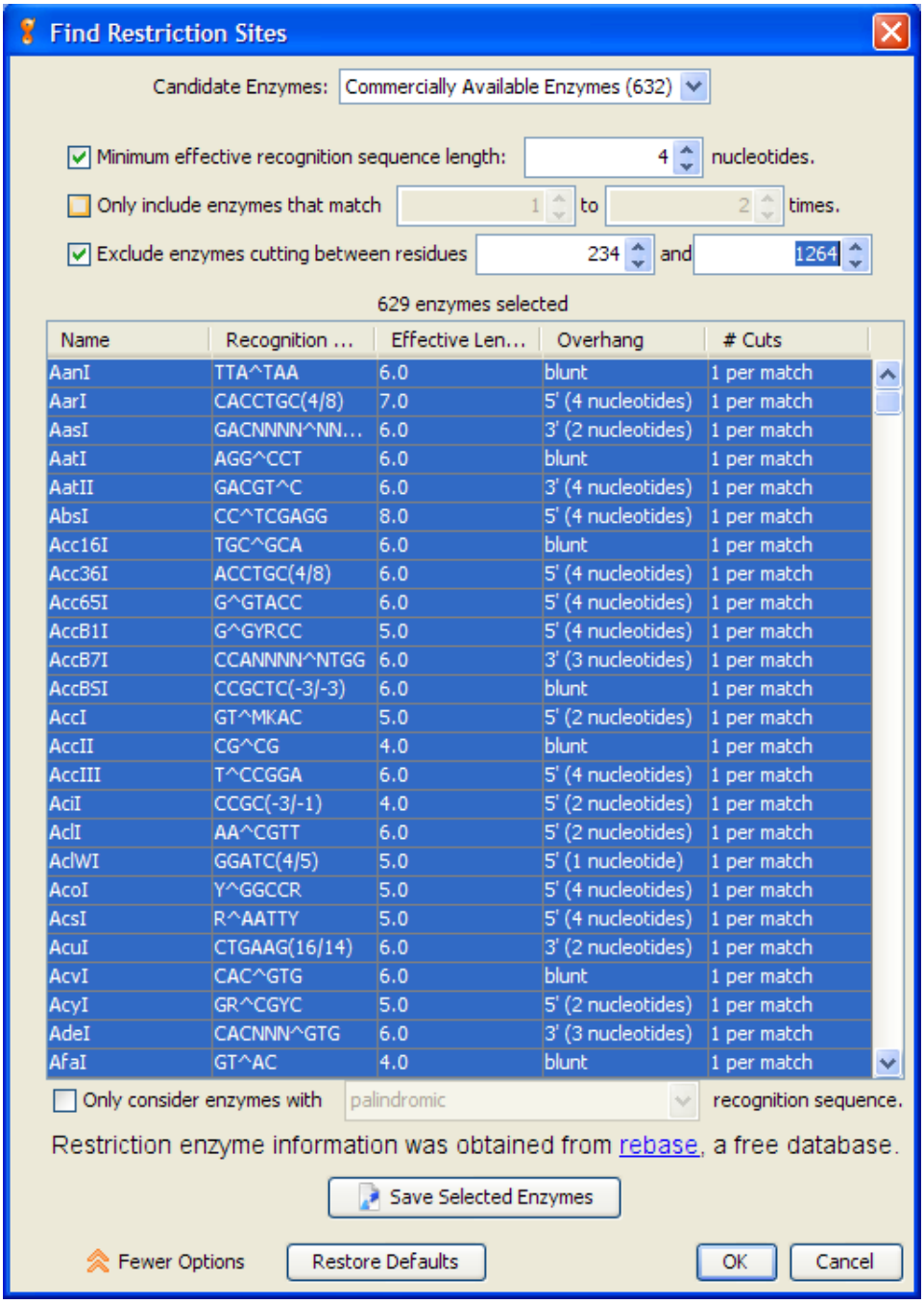

Find Restriction Sites
Candidate Enzymes: Commercially Available Enzymes (632)
Minimum effective recognition sequence length: 4 nucleotides.
Only include enzymes that match 1 to 2 times.
Exclude enzymes cutting between residues 234 and 1264
629 enzymes selected
Name | Recognition ... | Effective Len... | Overhang | # Cuts
AanI | TTA^TAA | 6.0 | blunt | 1 per match
AarI | CACCTGC(4/8) | 7.0 | 5' (4 nucleotides) | 1 per match
AasI | GACNNNN^NN... | 6.0 | 3' (2 nucleotides) | 1 per match
AatI | AGG^CCT | 6.0 | blunt | 1 per match
AatII | GACGT^C | 6.0 | 3' (4 nucleotides) | 1 per match
AbsI | CC^TCGAGG | 8.0 | 5' (4 nucleotides) | 1 per match
Acc16I | TGC^GCA | 6.0 | blunt | 1 per match
Acc36I | ACCTGC(4/8) | 6.0 | 5' (4 nucleotides) | 1 per match
Acc65I | G^GTACC | 6.0 | 5' (4 nucleotides) | 1 per match
AccB1I | G^GYRCC | 5.0 | 5' (4 nucleotides) | 1 per match
AccB7I | CCANNNN^NTGG | 6.0 | 3' (3 nucleotides) | 1 per match
AccBSI | CCGCTC(-3/-3) | 6.0 | blunt | 1 per match
AccI | GT^MKAC | 5.0 | 5' (2 nucleotides) | 1 per match
AccII | CG^CG | 4.0 | blunt | 1 per match
AccIII | T^CCGGA | 6.0 | 5' (4 nucleotides) | 1 per match
AciI | CCGC(-3/-1) | 4.0 | 5' (2 nucleotides) | 1 per match
AclI | AA^CGTT | 6.0 | 5' (2 nucleotides) | 1 per match
AclWI | GGATC(4/5) | 5.0 | 5' (1 nucleotide) | 1 per match
AcoI | Y^GGCCR | 5.0 | 5' (4 nucleotides) | 1 per match
AcsI | R^AATTY | 5.0 | 5' (4 nucleotides) | 1 per match
AcuI | CTGAAG(16/14) | 6.0 | 3' (2 nucleotides) | 1 per match
AcvI | CAC^GTG | 6.0 | blunt | 1 per match
AcyI | GR^CGYC | 5.0 | 5' (2 nucleotides) | 1 per match
AdeI | CACNNN^GTG | 6.0 | 3' (3 nucleotides) | 1 per match
AfaI | GT^AC | 4.0 | blunt | 1 per match
Only consider enzymes with palindromic recognition sequence.
Restriction enzyme information was obtained from rebase, a free database.
Save Selected Enzymes
Fewer Options
Restore Defaults
OK
Cancel

Figure 11.1: *Find Restriction Sites* options dialog, with extended options showing.

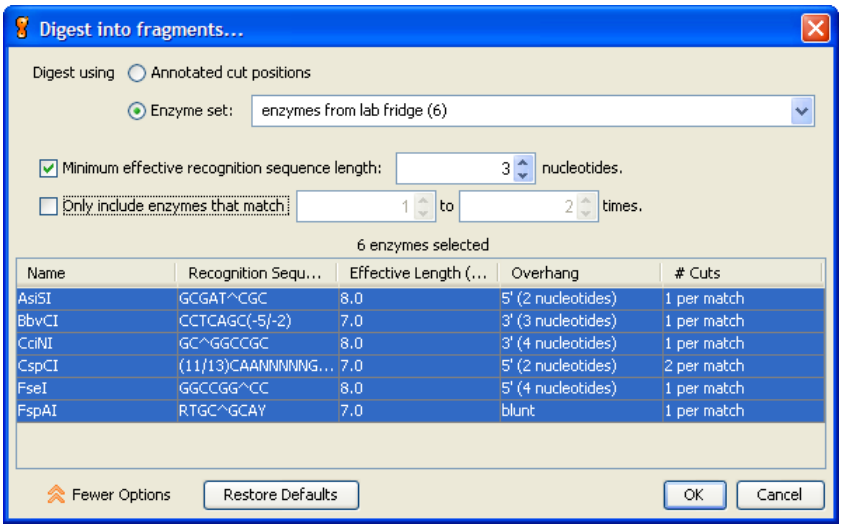


Figure 11.2: *Digest into fragments* options dialog, with extended options showing.

11.3 Insert into Vector

The option *Insert into Vector...* from the Tools / Restriction Analysis menu or the context menu allows you to take an insert and insert it into a vector. The insert must be one of the following:

- A fragment which has already been digested. This fragment cannot have any restriction site annotations on it. The entire fragment will be inserted into the vector. Overhangs will be taken into account.

- A sequence with two restriction annotations. The fragment resulting from digesting this sequence (and discarding the fragments from the ends) will be inserted into the vector.

The vector must be a circular sequence. You do not need to annotate the restriction sites used to cut the vector in advance; the Insert into Vector operation will do that for you.

This operation cannot deal with some aspects of molecular cloning such as triple ligation and the blunting or filling in of overhangs. If you want to do a cloning operation outside the scope of this operation, you will need to annotate restriction sites on the sequences involved, digest the fragments, modify them in the sequence viewer if necessary and then ligate them back together as a set of discrete steps.

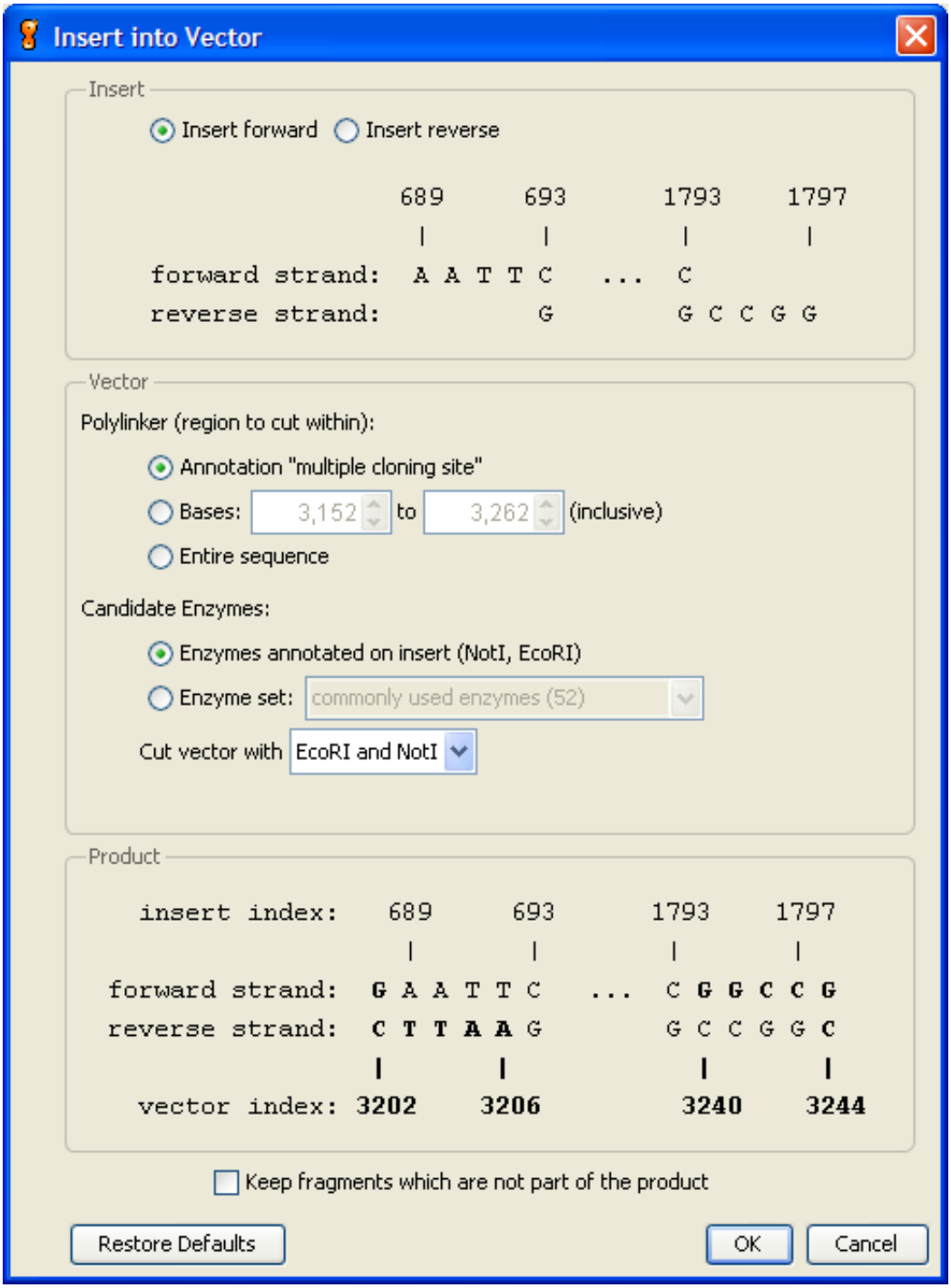

Insert into Vector
Insert
Insert forward
Insert reverse
689 693 1793 1797
forward strand: A A T T C ... C
reverse strand: G G C C G G
Vector
Polylinker (region to cut within):
Annotation "multiple cloning site"
Bases: 3,152 to 3,262 (inclusive)
Entire sequence
Candidate Enzymes:
Enzymes annotated on insert (NotI, EcoRI)
Enzyme set: commonly used enzymes (52)
Cut vector with EcoRI and NotI
Product
insert index: 689 693 1793 1797
forward strand: G A A T T C ... C G G C C G
reverse strand: C T T A A G G C C G G C
vector index: 3202 3206 3240 3244
Keep fragments which are not part of the product
Restore Defaults
OK
Cancel

Figure 11.3: *Insert into Vector* options dialog

11.3.1 Insert Options

You cannot alter the insert used in the operation from the options, but you can select what direction to insert in: forward or reverse. If the insert fragment has complementary overhangs or is blunt at both ends, you can also choose to insert in both directions. In this case, two product documents will be created, one for the insert in each direction.

The insert options also present a diagram showing the bases at each end of the insert fragment.

11.3.2 Vector Options

- *Polylinker (region to cut within):* These options let you choose what region within the vector sequence to look for enzymes to cut within. Geneious will examine the vector sequence for enzymes that have cut sites within this region and none outside it. You can specify the polylinker in the following ways:
  - *Annotation* If the vector has one or more polylinker annotations annotated on it, you can choose to use the interval covered by one such polylinker annotation directly.
  - *Bases* Used to explicitly specify the range of bases to use.
  - *Entire sequence* Used to specify that you can cut anywhere within the sequence.
- *Candidate Enzymes:* These options let you choose which enzymes to look for on the vector sequence
  - *Enzymes annotated on insert* This option lets you use only the enzymes used to cut the insert fragment.
  - *Enzyme set* This option lets you use the enzymes from a predefined enzyme set, eg. the enzyme set you have created containing the enzymes you have in your lab.
- *Cut vector with:* Whenever you change the options for the polylinker or candidate enzymes, Geneious will recalculate the compatible enzymes on the vector. It will look for enzymes which meet one of the following conditions (in addition to cutting only within the polylinker and belonging to enzymes from the candidate enzyme set):
  1. A single enzyme which cuts the vector once, such that the insert can be inserted in the gap (Possible only when the insert has complementary cut sites).
  2. A single enzyme which cuts the vector twice, such that the insert can be inserted into the gap vacated by the fragment between the two cut sites
  3. Two enzymes which each cut the vector once, such that the insert can be inserted into the gap vacated by the fragment between the two cut sites

11.3.3 Other Options

The Product section of the options displays a diagram showing the ligation points in the insertion. The parts of the ligation points belonging to the vector appear in bold in this diagram.

Below this is a checkbox where you can choose whether to *Keep fragments which are not part of the product*. If this box is checked, a document will be created representing the fragment removed from the vector, if any. If the insert fragment was produced from a sequence with two restriction site annotations, the fragments on either side of the restriction site annotations will also be kept.

11.4 Gateway® Cloning

Geneious contains three operations to assist with Gateway® cloning. Gateway is a registered trademark of Invitrogen Corporation.

11.4.1 Add AttB Sites

This operation allows you to add AttB sites to a PCR product. It will work on the following types of document:

- A PCR product. AttB sites will be appended to the PCR product.
- A document with primer binding sites annotated. If there is more than one pair, Geneious will ask you which pair to use. The PCR product will be extracted and AttB sites appended.
- Any other nucleotide document. The AttB operation will perform primer design on the document, then extract the PCR product and add AttB sites.

11.4.2 BP Reaction

This operation will create an entry clone from a donor vector (Vector with AttP sites) and either an expression clone (Vector with AttB sites) or a PCR product with AttB sites, as produced by the *Add AttB Sites* reaction. The resulting entry clone will have AttL sites annotated.

11.4.3 LR Reaction

This operation will create an expression clone from a destination vector (Vector with AttR sites) and an entry clone (Vector with AttL sites). The resulting entry clone will have AttB sites

annotated.

Chapter 12

Server Databases (*Pro* only)

By using server databases Geneious can store your documents in your favorite relational (SQL) database rather than on the file system. This means that multiple users can concurrently use the same synchronized storage location without any problems.

A server databases can be used for everything a local database is used for. This includes collaboration and smart agents. Take note that unread status, agents and shared folders belong to individual users rather than the database. For example Bob may see a document as unread, but Joe will see that same document as read if he has read it.

12.1 Supported Database Systems

To use a database as a server database Geneious requires that it support transactions with an isloation level set to SERIALIZABLE. Supported databases systems include Microsoft SQL Server, PostgreSQL, Oracle and MySQL. It is possible to use other database systems if you provide the database driver, see section 12.2.1

Server Databases have been tested using:

- Microsoft SQL Server 2005 Express
- PostgreSQL 7.4
- Oracle 10g Express Edition
- MySQL 5

12.2 Setting up

After a database is set up correctly, multiple users can connect to it and use it as their storage location just as if they were using their own local database.

Follow these steps to set up your database for use with Geneious.

- Install a supported database management system if you do not already have one.
- Create a new database with your desired name. Make sure that you have a user that has rights to create tables.
- Use the "Connect to a database button" to connect to your database. If the database has not been set up (usually the case if you are following these instructions) Geneious will detect this and set up the database. This will only succeed if you have permission to create tables on the database.
- Make sure any other users of the database have SELECT, INSERT, UPDATE and DELETE rights, otherwise they will not be able to use the Server Database as intended.

There are two ways you can use your database with multiple users. The simple way is just to use the server database as a shared local database. If this is all you want then you are now done with setup.

Alternatively you may want to restrict access to particular folders with groups and roles. To do this please refer to section 12.4.1.

Your database should now be ready to use with Geneious. Now all users can connect to the database by clicking on Server Databases in the service tree and then clicking "Connect to a setup database". This will bring up a dialog for the user to enter in the database details.

12.2.1 Supplying your own Database Driver

Server databases were designed with the supported databases in mind and packaged with database drivers for them. However Geneious allows you to supply your own jdbc database driver if you want to.

You may want to do this because you have an updated driver or because you have a driver for an unsupported database. It is not guarnteed that Server Databases will work with another database system if you provide its driver, but it is likely that it will.

To supply your own driver open up the dialog you would normally use to connect to a database. Then click the "More Options" button.

12.3 Removing a server database

To remove a server database, simply right click on its top folder and choose "Remove database".

12.4 Administration

The typical user will not have to do any administration, this section is for those in charge of the database.

12.4.1 Groups and Roles

Server databases support user groups and roles for managing access to documents. This means that you can restrict access of folders to privelleged people. How it works is that each folder in Geneious belongs to a group. Users can belong to any number of groups and have a specified role within that group. The three roles are:

- "View" allows the user to view the contents of folders.
- "Edit" allows the user to view and edit the contents of folders.
- "Admin" allows the user special administrative functions on folders.

As of this time Geneious only uses the "Admin" role for the "Everybody" group.

By default there is only one group, the "Everybody" group. When a user logs in for the first time Geneious will put them into the "Everybody" group with a role of "Edit". So this means every user of the server database belongs to this group with a role of "Edit" unless you enter them into the "g_user" table beforehand. You will want to give yourself the role of "Admin" for the "Everybody" group if you want to perform administrative functions within Geneious.

Unfortunately at this time there is no interface for assigning groups and roles to users. So you will need some knowledge of SQL in order to take advantage of this feature. You can create groups by adding entries into the "g_group" table in the database. Assign users groups and roles in the table "g_user_group_role".

It is likely that if you are running in a multi user environment and taking advantage of groups and roles you will want to give only read-access of the table "g_user_group_role" to your users. This is so your users can not edit this table with SQL directly as you would do. You will also want to add all of your users into "g_user" manually so Geneious does not think that they are first time users and fail trying to insert them into the "Everybody" group due to read-only access.

12.4.2 Database Indexing

Geneious indexes every document that is added to a server database for searching. It is very unlikely that this index will become corrupted. But if you are not getting correct search results or if you simply believe the database index has become corrupt somehow, the admin of the Everybody group can right click on the top folder of a server database to re-index it. This will not affect any other users until it is complete, however if your database contains many documents it will take a long time. Geneious must be left open to re-index the database.

Bibliography

[1] SF. Altschul, W. Gish, W. Miller, EW. Myers, and DJ. Lipman, *Basic local alignment search tool.*, J Mol Biol **215** (1990), no. 3, 403–410. 18, 26, 36

[2] MO. Dayhoff (ed.), *Atlas of protein sequence and structure*, vol. 5, National biomedical research foundation Washington DC, 1978. 80

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*, Cambridge University Press, 1998. 82

[4] J. Felsenstein, *Confidence limits on phylogenies: An approach using the bootstrap.*, Evolution **39** (1985), no. 4, 783–791. 88

[5] DF. Feng and RF. Doolittle, *Progressive sequence alignment as a prerequisite to correct phylogenetic trees.*, J Mol Evol **25** (1987), no. 4, 351–60. 83

[6] O. Gotoh, *An improved algorithm for matching biological sequences.*, J Mol Biol **162** (1982), 705–708. 80

[7] S. Guindon and O. Gascuel, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.*, Syst Biol **52** (2003), no. 5, 696–704. 86

[8] M. Vingron HA. Schmidt, K. Strimmer and A. von Haeseler, *Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing.*, Bioinformatics **18** (2002), no. 3, 502–504. 28

[9] M. Hasegawa, H. Kishino, and T. Yano, *Dating of the human-ape splitting by a molecular clock of mitochondrial dna.*, J Mol Evol **22** (1985), no. 2, 160–174. 88

[10] S. Henikoff and JG. Henikoff, *Amino acid substitution matrices from protein blocks.*, Proc Natl Acad Sci U S A **89** (1992), no. 22, 10915–10919. 80

[11] T. Jukes and C. Cantor, *Evolution of protein molecules*, pp. 21–32, Academic Press, New York, 1969. 88

[12] S. Kumar, K. Tamura, and M. Nei, *Mega3: Integrated software for molecular evolutionary genetics analysis and sequence alignment.*, Brief Bioinform **5** (2004), no. 2, 150–163. 31

[13] DR. Maddison, DL. Swofford, and WP. Maddison, *Nexus: an extensible file format for systematic information.*, Syst Biol **46** (1997), no. 4, 590–621. 28, 31, 87

[14] JV. Maizel and RP. Lenk, *Enhanced graphic matrix analysis of nucleic acid and protein sequences.*, Proc Natl Acad Sci U S A **78** (1981), no. 12, 7665–9. 78, 79

[15] C. Michener and R. Sokal, *A quantitative approach to a problem in classification.*, Evolution **11** (1957), 130–162. 86, 87, 91

[16] SB. Needleman and CD. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.*, J Mol Biol **48** (1970), no. 3, 443–53. 79, 80, 86, 103

[17] C. Notredame, DG. Higgins, and J. Heringa, *T-coffee: A novel method for fast and accurate multiple sequence alignment.*, J Mol Biol **302** (2000), no. 1, 205–217. 26

[18] RJ. Roberts, T. Vincze, J. Posfai, and D. Macelis, *Rebase – enzymes and genes for dna restriction and modification.*, Nucl Acids Res **35** (2007), D269–D270. 133

[19] F. Ronquist and JP. Huelsenbeck, *Mrbayes 3: Bayesian phylogenetic inference under mixed models.*, Bioinformatics **19** (2003), no. 12, 1572–4. 86

[20] N. Saitou and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.*, Mol Biol Evol **4** (1987), no. 4, 406–25. 86, 87, 91

[21] TF. Smith and MS. Waterman, *Identification of common molecular subsequences*, Journal of Molecular Biology **147** (1981), 195–197. 79, 80

[22] K. Tamura and M. Nei, *Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees.*, Mol Biol Evol **10** (1993), no. 3, 512–526. 88

[23] JD. Thompson, TJ. Gibson, F. Plewniak, F. Jeanmougin, and DG. Higgins, *The clustal x windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.*, Nucleic Acids Res **25** (1997), no. 24, 4876–4882. 24, 26, 28, 83, 84

[24] JD. Thompson, DG. Higgins, and TJ. Gibson, *Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.*, Nucleic Acids Res **22** (1994), no. 22, 4673–4680. 24, 28, 84